THE JOURNAL OF
DUCATIONAL PSYCHOLOGY

me XXXI May, 1940 Number 5

CTOR ANALYSIS IN THE ESTABLISHMENT OF NEW PERSONALITY TESTS[1]

JOHN G. DARLEY

University of Minnesota

AND

WALTER J. McNAMARA

International Business Machines Corporation

In an earlier article,[2] the present authors described the application Thurstone's method of factor analysis to both test and retest ormance on thirteen attitude and adjustment scales. The average rval between testing and retesting was nine and two-tenths months. ough certain theoretical questions were raised regarding the effect hanges in measured behavior on the factor loadings, the data ed sufficiently stable to warrant further study leading to the blishment of new and more homogeneous scales for the measure- t of the behavior constellations identifiable in the factor analysis. s report presents the method used in setting up experimental forms hese new personality measures.

Table I lists the factor loadings in the test and retest situation ved from intercorrelations for one hundred men. Table II lists ilar data for one hundred women. These tables have been rear- ged from tables in the earlier article to provide a uniform order of ors in both test and retest analysis.

Assistance in the preparation of the item analyses contained in this article furnished by the personnel of the Works Progress Administration Official ct Number 665-71-3-69.

McNamara, W. J. and Darley, J. G.: "A Factor Analysis of Test-Retest rmance on Attitude and Adjustment Tests." *Journal of Educational Psy- gy*, Vol. XXIX, No. 9, December, 1938, pp. 652–664.

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

olume XXXI | May, 1940 | Number 5

ACTOR ANALYSIS IN THE ESTABLISHMENT OF NEW PERSONALITY TESTS[1]

JOHN G. DARLEY

University of Minnesota

AND

WALTER J. McNAMARA

International Business Machines Corporation

In an earlier article,[2] the present authors described the application Thurstone's method of factor analysis to both test and retest rformance on thirteen attitude and adjustment scales. The average erval between testing and retesting was nine and two-tenths months. though certain theoretical questions were raised regarding the effect changes in measured behavior on the factor loadings, the data med sufficiently stable to warrant further study leading to the ablishment of new and more homogeneous scales for the measure- nt of the behavior constellations identifiable in the factor analysis. is report presents the method used in setting up experimental forms these new personality measures.

Table I lists the factor loadings in the test and retest situation rived from intercorrelations for one hundred men. Table II lists nilar data for one hundred women. These tables have been rear- nged from tables in the earlier article to provide a uniform order of tors in both test and retest analysis.

---

[1] Assistance in the preparation of the item analyses contained in this article furnished by the personnel of the Works Progress Administration Official oject Number 665-71-3-69.

[2] McNamara, W. J. and Darley, J. G.: "A Factor Analysis of Test-Retest rformance on Attitude and Adjustment Tests." *Journal of Educational Psy- logy*, Vol. XXIX, No. 9, December, 1938, pp. 652–664.

TABLE I.—LOADINGS OF FIVE FACTORS IN THIRTEEN ATTITUDE AND ADJUSTMENT SCALES, BASED ON FACTOR ANALYSIS OF TEST AND RETEST INTERCORRELATIONS FOR ONE HUNDRED MEN

Scale	Factor				
	I	II	III	IV	V
Test					
Minnesota scale for survey of opinions:					
Morale	.680	.379	.028	.187	.115
Inferiority	.338	.680	.012	.011	.027
Family	.450	.032	.689	−.032	−.118
Law	.618	.100	.069	.148	.126
Economic conservatism	.129	.059	−.120	.145	.624
Education	.700	−.089	.027	−.051	−.003
General adjustment	.814	.274	−.014	.216	.332
Adjustment inventory:					
Home	.151	.131	.467	.454	.005
Health	−.033	−.017	.334	.460	.009
Social	−.065	.813	.232	.210	.164
Emotional	.107	.315	.154	.745	.070
Minnesota inventory of social attitudes:					
Social preferences	−.044	.327	.514	.032	.341
Social behavior	.008	.695	.307	.017	.256
Retest					
Minnesota scale for survey of opinions:					
Morale	.727	.269	.134	.202	.262
Inferiority	.380	.444	.150	.481	.230
Family	.389	.126	.644	−.062	.028
Law	.629	.035	.233	.040	.253
Economic conservatism	−.034	.063	.012	−.108	.791
Education	.682	.202	−.103	−.074	−.060
General adjustment	.765	.262	.025	.049	.441
Adjustment inventory:					
Home	.133	.055	.762	.262	.150
Health	−.121	−.086	.256	.476	−.009
Social	.047	.636	.004	.543	.143
Emotional	.047	.095	.361	.739	.244
Minnesota inventory of social attitudes:					
Social preferences	.064	.742	.220	.051	.021
Social behavior	.062	.840	.088	.378	.047

TABLE II.—LOADINGS OF FIVE FACTORS IN THIRTEEN ATTITUDE AND ADJUSTMENT SCALES, BASED ON FACTOR ANALYSIS OF TEST AND RETEST INTERCORRELATIONS FOR ONE HUNDRED WOMEN

Scale	Factor				
	I	II	III	IV	V
Test					
Minnesota scale for survey of opinions:					
Morale	.870	.119	.038	.156	.226
Inferiority	.520	.486	.056	.238	.113
Family	.586	.076	.606	−.056	.112
Law	.672	−.082	.019	.160	.520
Economic conservatism	.281	.094	.034	−.027	.557
Education	.710	−.013	−.114	.076	.027
General adjustment	.878	.040	.100	.128	.187
Adjustment inventory:					
Home	.203	.164	.678	.386	.110
Health	−.094	.045	.249	.534	.145
Social	.199	.786	.068	.120	−.112
Emotional	.205	.335	.220	.572	−.042
Minnesota inventory of social attitudes:					
Social preferences	.010	.593	−.095	.035	.373
Social behavior	.066	.867	.078	−.023	.112
Retest					
Minnesota scale for survey of opinions:					
Morale	.830	.136	.222	.126	.346
Inferiority	.433	.500	.151	.336	.218
Family	.232	.040	.720	.416	.048
Law	.571	.003	.381	−.081	.308
Economic conservatism	−.009	.103	.645	−.121	−.002
Education	.666	−.105	.035	.021	.036
General adjustment	.798	.165	.500	.057	−.031
Adjustment inventory:					
Home	.042	.017	.498	.713	−.039
Health	−.035	−.058	−.111	.621	.229
Social	−.069	.744	.032	.358	.105
Emotional	.227	.204	.098	.804	.026
Minnesota inventory of social attitudes:					
Social preferences	.067	.312	.161	.061	.528
Social behavior	.141	.854	.116	.230	.170

The Minnesota Scale for the Survey of Opinions[1] contains one hundred thirty-two items arranged in serial order to measure six areas of behavior. Thus there are twenty-two items in each of the following scales: Morale; inferiority; family adjustment; attitude toward law; economic conservatism; attitude toward education. A seventh scale—general adjustment—is derived by regrouping and separate scoring of sixteen items contained within the original number of one hundred thirty-two items.

All items are phrased impersonally, with five possible response positions, as follows:

HOME IS THE MOST PLEASANT PLACE IN THE WORLD.
Strongly agree. Agree. Undecided. Disagree. Strongly disagree.

The Adjustment Inventory[2] contains one hundred forty items arranged in random order to measure four areas of behavior. There are thirty-five items in each of the following scales: Home adjustment; health adjustment; social adjustment; emotional adjustment.

All items are phrased as questions in the second person singular, with three possible response positions, as follows:

YES. NO. ?. Do you day-dream frequently?

The Minnesota Inventories of Social Attitudes[3] comprise two scales of forty items each. Form *P* (Social Preferences) is designed to measure the individual's preferences for gregarious or more solitary social activities. Form *B* (Social Behavior) is designed to measure the individual's self-estimates of his own skills and ease in social situations.

All items are phrased in the third person singular with five possible response positions, as follows:

(Form *P*)

LIKES TO MIX WITH PEOPLE SOCIALLY.
Almost always. Frequently. Occasionally. Rarely. Almost never.

(Form *B*)

IS RELUCTANT TO MEET IMPORTANT PEOPLE.
Almost always. Frequently. Occasionally. Rarely. Almost never.

---

[1] Rundquist, E. A. and Sletto, R. F.: *Scoring Instructions for the Minnesota Scale for the Survey of Opinions.* University of Minnesota Press, 1936.

[2] Bell, H. M.: *Manual for the Adjustment Inventory.* Stanford University Press, 1934.

[3] Williamson, E. G. and Darley, J. G.: *Manual for the Minnesota Inventories of Social Attitudes.* The Psychological Corporation, New York City, 1937.

An inspection of Tables I and II reveals certain logical groupings of these thirteen scales that may be derived from the factor loadings. In the absence of sampling data on the distribution and critical levels of factor loadings, one is thrown back upon empiric considerations in determining either statistical significance or psychological meaning. In these tables, the problem was first to regroup the thirteen scales in such a way that all would be accounted for within five factors, that no one scale would appear in more than one factor, and that the groupings would be the same for both test and retest, and men and women. Once such a grouping is established, a summated factor score can be set up for item analysis. These five new factor scores may include scales whose present labels would seem to set them apart, but the logic structure of factor analysis should then lead to a reëxamination of the content of the scales to determine underlying behavioral similarities through which their loading of a given factor becomes psychologically meaningful.

With these empiric considerations in mind, the scales were distributed to the five factors as follows:

Factor I includes: Morale; attitude toward law; attitude toward education; general adjustment.

Factor II includes: Inferiority; social adjustment; social preferences; social behavior.

Factor III includes: Family adjustment; home adjustment.

Factor IV includes: Health adjustment; emotional adjustment.

Factor V includes: Economic conservatism.

These groupings, for test and retest performance as well as for men and women, represent the best evaluation of factor loadings that could be worked out without including a given scale in more than one factor group. They also have rudimentary meaning deriving from the present labels of the scales. Thus Factor I seems to comprise a measure of adjustment to society in general as well as some of its institutions. Factor II appears to measure the individual's effectiveness in face-to-face social situations. Factor III is specifically related to family adjustment. Factor IV might be relabelled as neurotic tendencies, if the label were in better repute. Factor V is apparently a specific aspect of radicalism-conservatism.

The next step was to convert all scores on the thirteen scales to sigma scores by use of the standard formula:

$$\frac{x - \bar{x}}{\sigma_x}$$

This conversion gave a set of relative scores that could be handled additively, independent of the score range or scoring methods found in the original scales. Five factor scores on both testing and retesting were then computed for each individual in the samples of men and women, to replace the thirteen original scores obtained in both the test and retest situation. The factor scores were obtained by adding the sigma scores for the scales as they were regrouped within each factor.

Table III gives the inter-correlations of the new factor scores for men and women separately on test and retest situations. Even without applying item analysis methods to the data to increase the homogeneity of the new factor scales, these intercorrelations indicate a fair degree of independence of the five factor scores. It is also noteworthy that the average intercorrelation in this table is only .25, lower than the average intercorrelations reported for the thirteen scales from which the factor groups were derived.

TABLE III.—INTERCORRELATIONS OF FIVE NEW FACTOR SCORES DERIVED BY REGROUPING AND ADDING SIGMA SCORES ON ORIGINAL THIRTEEN SCALES

	One hundred men original test	One hundred men retest	One hundred women original test	One hundred women retest
Factor I *vs.* Factor II	.29	.39	.30	.32
Factor I *vs.* Factor III	.39	.37	.45	.39
Factor I *vs.* Factor IV	.19	.10	.18	.18
Factor I *vs.* Factor V	.27	.28	.39	.20
Factor II *vs.* Factor III	.32	.28	.29	.32
Factor II *vs.* Factor IV	.33	.41	.31	.34
Factor II *vs.* Factor V	.23	.11	.22	.09
Factor III *vs.* Factor IV	.40	.38	.41	.50
Factor III *vs.* Factor V	−.05	.04	.16	.18
Factor IV *vs.* Factor V	.13	−.01	.05	−.14

The factor scores lead naturally to the possibility of item analysis. For this purpose the names of the twenty-five highest and the twenty-five lowest cases within each factor score distribution were traced for use as criterion groups. At this point a further reference to Tables I and II is necessary. It will be seen that the empiric grouping of the thirteen original scales forced certain scales into a given factor even

though those scales had fairly high loadings in one or more other factors as well. For example, the family scale for men and women was assigned to Factor III; yet it has loadings of .450 and .389, on test and retest, respectively, in Factor I for men, and loadings of .586 and .232 on test and retest, respectively, in Factor I for women. Other examples of this nature may also be seen in Tables I and II. It is conceivable that a scale could have significant loadings in more than one factor; the items in the scale may each have overlapping contributions or they may subdivide in such a way that groups of items within the scale carry relatively exclusive loadings of more than one factor. To study this phenomenon, the item analyses were carried out in each factor on more than the basic number of scales that went in to the new factor score. The actual scales for which item analyses were made were as follows:

	Men		Women	
	Test	Retest	Test	Retest
Factor I.....	Morale	Morale	Morale	Morale
	Inferiority	Inferiority	Inferiority	Inferiority
	Family adjustment	Family adjustment	Family adjustment	Family adjustment
	Law	Law	Law	Law
	Education	Education	Education	Education
	General adjustment	General adjustment	General adjustment	General adjustment
Factor II....	Morale	Morale		
	Inferiority	Inferiority	Inferiority	Inferiority
	General adjustment	General adjustment		
	Social adjustment	Social adjustment	Social adjustment	Social adjustment
	.....	.....	Emotional adjustment	Emotional adjustment
	Social preference	Social preference	Social preference	Social preference
	Social behavior	Social behavior	Social behavior	Social behavior
Factor III...	Family adjustment	Family adjustment	Family adjustment	Family adjustment
	Home adjustment	Home adjustment	Home adjustment	Home adjustment
	Health adjustment	Health adjustment		
	Social preference	Social preference		
Factor IV....	.....	.....	Inferiority	Inferiority
	Home adjustment	Home adjustment	Home adjustment	Home adjustment
	Health adjustment	Health adjustment	Health adjustment	Health adjustment
	Social adjustment	Social adjustment		
	Emotional adjustment	Emotional adjustment	Emotional adjustment	Emotional adjustment
Factor V.....	.....	.....	Morale	Morale
	Law	Law	Law	Law
	Economic conservatism	Economic conservatism	Economic conservatism	Economic conservatism
	General adjustment	General adjustment		
	.....	.....	Social preference	Social preference

Admittedly, when these additional scales are included for item analysis in the high-low groups within the five new factor scores, the factor scores tend to lose some of their earlier meaning. On the other hand, the factor loadings of these additional scales, as seen in Tables I and II, cannot be disregarded in an attempt to isolate homogeneous items. For this latter reason the item analyses included more scales than appeared in the groupings whose sigma scores were added to give the factor scores. However, the criterion groups were selected solely on the basis of the scales grouped in the factors as given on page 327.

To foreshadow the results of the item analyses, it may be well to summarize the total number of items within each factor for which critical ratios were established:

	Men		Women	
	Test	Retest	Test	Retest
Factor I	126	126	126	126
Factor II	175	175	172	172
Factor III	132	132	57	57
Factor IV	140	140	127	127
Factor V	60	60	106	106

In the case of the general adjustment scale (sixteen items) these figures represent a slight overlapping, since the items in this scale appear also in the six main scales of the Minnesota Scale for the Survey of Opinions. This overlapping was later eliminated by assigning these items to their original scale location. In all the items analyzed for the new Factor I test, only two showed critical ratios of 3.00 or greater in a factor other than Factor I. Nine items having critical ratios of 3.00 or greater in Factor II also had critical ratios of 3.00 or greater in other factors. Nine items with critical ratios of 3.00 or greater in Factor III recurred in other factors with similar critical ratios. Five items with critical ratios of 3.00 or better in Factor IV also had high critical ratios in other factors. None of the items with critical ratios of 3.00 or greater in Factor V appeared in the other factors with equally high critical ratios. Thus, only twenty-five items showed overlapping that produced critical ratios of 3.00 or more in two different factors. Since twenty-four hundred forty-two critical ratios were derived in this item analysis, there is approximately one per cent of overlapping as here

defined. An inspection of the critical ratios for these twenty-five items was made to determine where the item should finally be located. Critical ratios of 3.00 or better for both sexes in both testing and retesting were given preference over critical ratios of the same size occurring for only one sex in both testing and retesting. In this manner all twenty-five items were located within the five factors without overlap.

With all critical ratios established for each sex separately and for both test and retest situations, the next step involved the selection of items which seemed to carry the greatest weight in the new factor scores, regardless of the original scale from which they were drawn. Items that differentiated the top twenty-five cases and the bottom twenty-five cases to the extent of a critical ratio of 3.00 or more were chosen first. Furthermore, only such items as provided this critical ratio in both testing and retesting were chosen. And finally, since the original thirteen scales all contained sex differences of varying magnitudes, the items were grouped according to their differential power for both sexes, for men separately and for women separately. This latter division permits the standardization of new tests for each sex, a much-needed technique in personality measurement.

Table IV indicates the number of items so selected from each original scale for inclusion in each new factor test; the table also indicates the number of items with critical ratios of 3.00 for both sexes, for men alone, and for women alone, in both testing and retesting. Table IV contains no overlapping.

It is first apparent in Table IV that the scales whose sigma scores were added to give the new factor score contribute the greatest number of items to any new factor test. The scales not included in this additive process, but included in the item analysis because of possibly significant factor loadings, contribute relatively few items to a new factor test. For example, the items in the inferiority scale were analyzed in relation to the new high and low scores of Factor I because this scale had substantial loadings of Factor I. Yet it contributes one item for both sexes, two other items for men alone, and no items for women alone that show critical ratios of 3.00 or greater. The items of the inferiority scale were also analyzed for women in relation to the new additive scores in Factor IV, and none of the items differentiated the high and low groups of women in this test. Yet the scale does contribute a total of fifteen items of differential value in the group of scales under Factor II, where it was included in the additive process to get the Factor II score.

TABLE IV.—NUMBER OF ITEMS FROM EACH ORIGINAL SCALE WITH CRITICAL RATIOS OF 3.00 FOR BOTH SEXES, FOR MEN ONLY AND FOR WOMEN ONLY, WITHIN EACH NEW FACTOR SCORE

Original scale	Number of items	Number of items contributed to factor															Total
		I			II			III			IV			V			
		Both sexes	Men only	Women only	Both sexes	Men only	Women only	Both sexes	Men only	Women only	Both sexes	Men only	Women only	Both sexes	Men only	Women only	
Morale	22	7	2	3	..	—	..	..	..	..	..	..	..	..	..	—	12
Inferiority	22	1	2	—	6	2	7	..	..	..	..	..	—	..	..	..	18
Family	22	1	1	—	..	..	..	10	1	4	..	..	..	..	..	..	17
Law	22	4	6	4	..	..	..	..	..	..	..	..	..	—	—	—	14
Economic conservatism	22	..	..	..	..	..	..	..	..	..	..	..	..	12	4	1	17
Education	22	3	2	6	..	..	..	..	..	..	..	..	..	..	..	..	11
Home	35	..	..	..	..	..	..	7	2	10	—	—	2	..	..	..	21
Health	35	..	..	..	..	..	..	..	—	..	1	4	12	..	..	..	17
Social	35	..	..	..	9	15	1	..	..	..	..	—	..	..	..	..	25
Emotional	35	..	..	..	..	..	—	..	..	..	13	3	4	..	..	..	20
Social Preferences	40	..	..	..	8	9	7	..	—	..	..	..	..	..	..	—	24
Social Behavior	40	..	..	..	20	5	2	..	..	..	..	..	..	..	..	..	27
Total	352	16	13	13	43	31	17	17	3	14	14	7	18	12	4	1	223

Note.—A dash (—) under any column indicates that the scale against which it appears was analyzed but contributed no items with critical ratios of 3.00 or better.

Since Table IV lists the number of items in each original scale in parentheses after the scale name, it is possible to observe the proportion of such items that yield critical ratios of 3.00 or better in the item analyses within the new factor scores. Sex differences are also noticeable in the table; for example, the social adjustment scale contributes to the Factor II test nine items of differential value for both sexes, fifteen items of differential value for men only, and one such item for women only.

Since only twenty-five cases appeared in the high- and low-score groups for item analysis purposes, it is possible that a critical ratio of 3.00 or greater might eliminate some items of value from the experimental forms of the five new factor tests. Therefore, Table V indicates the additional items that appear when a critical ratio of 2.00 is set as the required differential limit. These additional items do not duplicate those already listed in Table IV.

The next step in the construction of trial forms of the new factor tests was to identify and edit to a common form the items sorted out in the item analyses. As was stated earlier, the original thirteen scales contained three different forms of statement and three differently-phrased response possibilities. The editing was designed to rephrase items to a common form where necessary and to establish consistent response possibilities within the new scales. Since all the items for the new Factor I test came originally from the Minnesota Scale for the Survey of Opinions, no editing was needed. Although all the items for the Factor IV test came originally from the Adjustment Inventory, it was judged best to set up five answer positions rather than the three positions originally used in the Adjustment Inventory. Slight modifications were, therefore, made in these items and instead of "Yes, No, ?" as possible answers, the phrases "Almost always, Frequently, Occasionally, Rarely, Almost never" were used as possible answers. The new test for Factor V again offered no problem of editing, but, since there were relatively few items in it, twenty-one items were adapted for experimental use from a scale recently established by Pace[1] for the measurement of economic attitudes.

The new tests for Factors II and III presented the most difficult problems in editing. In the first place, in cases of duplicating items only the most clearly stated form was retained. In the second place,

[1] Pace, C. R. Jr.: *The Relationship Between Liberal-Conservative Attitudes and Knowledge of Current Affairs*, July, 1937, Ph.D. thesis on file at the University of Minnesota Library.

TABLE V.—NUMBER OF ITEMS FROM EACH ORIGINAL SCALE WITH CRITICAL RATIOS OF 2.00 FOR BOTH SEXES, FOR MEN ONLY, AND FOR WOMEN ONLY, WITHIN EACH NEW FACTOR SCORE

Original scale	Number of items contributed to factor																			
	I			II			III			IV			V							
	Both sexes	Men only	Women only	Both sexes	Men only	Women only	Both sexes	Men only	Women only	Both sexes	Men only	Women only	Both sexes	Men only	Women only					
Morale	2	1	6	..	..	..	..	..	..	..	..	..	..	..	..					
Inferiority	..	..	..	2	1	1	..	..	..	..	..	..	..	..	..					
Family	..	..	..	..	..	..	2	1	2	..	..	..	..	..	..					
Law	—	—	5	..	..	..	..	..	..	..	..	..	..	..	..					
Economic conservatism	..	..	..	..	..	..	..	..	..	..	..	..	1	2	1					
Education	3	—	5	..	..	..	..	..	..	..	..	..	..	..	..					
Home	..	..	..	..	..	..	2	3	2	..	..	..	..	..	..					
Health	..	..	..	..	..	..	..	..	..	3	1	5	..	..	..					
Social	..	..	..	2	3	3	..	..	..	..	..	..	..	..	..					
Emotional	..	..	..	..	..	..	..	..	..	3	1	4	..	..	..					
Social preferences	..	..	..	4	2	2	..	1	..	..	..	..	..	..	..					
Social behavior	..	..	..	1	1	1	..	..	..	..	..	..	..	..	..					
Total	5	1	16	9	7	7	4	5	4	6	2	9	1	2	1					

items for these new tests came from original scales that included all three types of item and response position; no data existed to establish a preference for any one type of item or response position. Accordingly, a further study was made of two forms of a test for Factor III. In Form *A* the items were stated impersonally, as follows:

ONE FEELS MOST CONTENTED AT HOME.
Strongly agree. Agree. Undecided. Disagree. Strongly disagree.

In Form *B* the same items were phrased in the second person singular, as follows:

DO YOU FEEL MOST CONTENTED AT HOME?
Almost always. Frequently. Occasionally. Rarely. Almost never.

There were forty-five items in each form, differing only as indicated in the samples given above. The two trial forms were given to one hundred college sophomores in laboratory psychology, fifty men and fifty women. Half the group took Form *A* first, followed by Form *B* one week later; the other half took Form *B* first followed by Form *A* one week later. The scoring was so arranged that a high score represented better measured adjustments. Table VI summarizes the resulting statistics.

At the conclusion of the second testing, these students were asked to indicate which form of test they preferred. More than eighty students favored Form *B*—the personal form of statement. This was the form on which the scores were most favorable, as may be seen in Table VI.

TABLE VI.—SUMMARY OF STATISTICAL CONSTANTS DETERMINED FROM FORMS *A* AND *B* OF NEW TEST FOR FACTOR III

	Form *A*		Form *B*		Correlations between Forms *A* and *B*	Critical ratios Form *A* vs. *B*
	Mean	Sigma	Mean	Sigma		
Fifty men	155.04	24.5	172.44	26.4	.88	9.8
Fifty women	158.88	22.7	179.28	19.8	.73	9.1
One hundred men and women	156.96	23.7	175.86	23.6	.81	13.1

Although this analysis gave no clear-cut answer regarding the most effective form of statement, it at least indicated student preferences, and these preferences seemed sufficiently clear to deserve consideration.

Therefore, the trial forms of new tests for Factors II and III were cast in the second person singular, question form. They then corresponded with the phrasing of the Factor IV test.

The trial forms of the five new factor tests show the following phrasing:

Factor I—impersonal statements, calling for agreement or disagreement.
Factor II—personal questions, calling for estimations of frequency.
Factor III—same as Factor II.
Factor IV—same as Factor II.
Factor V—same as Factor I.

These trial forms are now being used preparatory to a final check on their internal consistency, meaning, and establishment of norms.

A word may be said in conclusion regarding the reliability and validity of these new tests. Theoretically, the use of factor analysis and item analysis techniques on these data should yield groupings of items of greater homogeneity than was found in the original thirteen scales; such items would then show higher odd-even correlations as one measure of reliability. From the standpoint of validity, factor analysis establishes one form of validity in terms of internal consistency and, furthermore, none of the validity of the original scales should be lost in this greater refinement of measurement. However, these theoretical possibilities can be most critically tested by applying the new factor tests to new populations. The results of the study of the trial forms will appear at a later date.

CONSTANCY AND VARIATION IN PATTERNS OF FACTOR LOADINGS

CHARLES M. HARSH

Randolph-Macon Woman's College

In the past few years many persons, sceptical of a new and unfamiliar device, have expressed doubt of the meaningfulness and dependability of factor analysis. On the one hand, it has been argued that the method is mathematically artificial, and too rigid to fit the facts of psychology. On the other hand, it has been objected that factorizations are not rigid enough, that too much is arbitrary, and that the reliability of factor loadings can not be calculated. Surely the interested sceptic deserves concrete evidence upon which to base his opinions, yet the arguments have usually been devoid of actual illustrative examples. Of the few examples given, some have been interpreted with misleading implications which need to be corrected. To complicate further the picture there are various schools of thought as to how factors should be interpreted. There are those who have faith that factors, when sufficiently purified, will ultimately represent fundamental unitary determiners of behavior. Others consider factorizations as convenient representations of relationships, to be presented with crossed fingers and alphabetical designations. Still others believe that factors should be calculated, sneered at, and thrown in the waste basket. And, finally, the conservatives conclude that factorizations are too laborious to be worth while anyway. The practical evaluation will depend not upon the correctness of any of these viewpoints, but rather upon the scientific utility of the results of factor studies.

Objections concerning the significance of mathematical factors have recently been summarized and conservatively answered by Garrett.[4] A short article by Thurstone[18] has pointed out that many of these objections arise from misapplication of factorial methods, in that factorization is not carried far enough, centroid axes are not rotated to conform to psychologically interpretable dimensions, or accepted criteria of adequate analysis are not met. For the benefit of those not familiar with the factor literature several of these points will be illustrated below, but the chief consideration of this paper will be the variability of factor loadings.

Concerning the dependability of factor analyses one can ask several questions: (1) How reliable is a given factor loading resulting from an

analysis of a table of intercorrelations? (2) Will a given test have the same common factor loadings when analyzed in combination with various groups of tests? Or, to restate the problem, will the same common factors be found, and recognized, when varying test batteries are used? (3) How variable is a given pattern of factor loadings for a set of measures on a sample of a certain population? For example, if the same set of tests is given to apparently similar population samples, and for each sample the tests are correlated and factorized, how much will the patterns of factor loadings vary? (4) Within what limits can a factor pattern for a given set of tests be expected to remain constant over a period of time between administrations of the test battery? (5) What variations in factor pattern should be expected in different groups of subjects?

In so far as possible these questions will be considered separately.

1. RELIABILITY OF A LOADING

The lack of an exact estimate of the probable error of a factor loading has been insistently deplored by Kelley.[7] No one will deny the desirability of such a measure of variability, but in its absence we may yet glean some indications from logical considerations and from observed variations. (Sometimes empirical justification is nearly as convincing as a mathematical derivation.) (*a*) In the first place, a factor loading is determined quite directly from a sum of several correlation coefficients, whose reliabilities depend upon their magnitude. Therefore, the reliability of a factor loading must likewise be related to the magnitude of the correlations from which it is calculated. (*b*) The reliability of the correlation coefficients also depends upon the size of the sample population; consequently, it seems sensible to conclude that the reliability of a factor loading depends upon the size of the population. But several coefficients are being pooled in estimating a loading, and one might expect errors in some coefficients to be counteracted by opposite errors in others, in which case one might expect factor loadings to be less variable than the individual correlations. (*c*) If this is so, then the reliability of a loading should depend upon the number of correlations from which it is calculated, and, consequently, upon the number of tests involved, in that the greater the number of coefficients pooled, the better the chance that errors in them will cancel. (*d*) Errors of the analytic procedure, especially the errors in estimating communalities, will have a relatively greater effect on the size of factor loadings when few tests

are involved. Thus, again, the reliability of loadings should be greater when more tests are involved. (*e*) Actually, the effects of errors in estimating communalities (h^2) are cumulative, in extracting successive factors. Consequently, the reliability of factor loadings must depend upon how many factors have already been extracted from the matrix. With poor estimates of h^2 and with small numbers of variables, these cumulative errors have often led to reports of impossibly high variances totaling more than 1.00, but with experience one can learn to make rather good estimates of h^2. (*f*) Finally, the reliability of a factor loading may depend upon the importance of the common factor, *i.e.*, the total variance in all tests accounted for by the factor. It has been reported that principal components beyond the first are not reliable,[8] but this may apply only in cases in which later components are very unimportant. McNemar* reports very consistent first centroid factor loadings for tests administered to several successive age groups in standardizing the new Stanford Binet. But his second factor loadings, which are so small as to be negligible up to the fourteen-year level, are quite inconsistent.

There is great need for a similar kind of evidence from analyses yielding larger loadings beyond the first factor. The present writer had occasion to administer an Annoyance Inventory to three populations of from two hundred to three hundred fifty college students. Scores on thirteen categories of annoyance were intercorrelated and factorized. The first centroid factor loadings were most consistent, and the loadings of successive factors became more variable as the factors decreased in importance. It was further observed that after rotating the axes to give simpler structure, the loadings of the second, third, and fourth simple factors were rather more consistent than the loadings of the second, third, and fourth centroid factors. As these simple factors had higher loadings than the corresponding centroid axes, it might be argued that the increased weighting led to increased consistency of factor loadings. But it might merely illustrate the point that centroid loadings do not reveal a factor structure as clearly or consistently as do the simple-factor loadings arrived at by rotation. Minor changes of correlations or different selection of tests for reflection may shift the balance and considerably alter the nature of the centroid factors without appreciably changing the general factor structure, (as will be illustrated in the next section). The present

* Remarks addressed to a round table on Factor Analysis at the meetings of the A.P.A. at Columbus, Ohio, September, 1938.

scarcity of evidence does not permit an estimation of the magnitude of each of the above-mentioned influences upon factor loading reliability, but it should be possible more expressly to design factor studies to reveal the magnitude of such influences.

One such study has just been reported by Mosier,[11] who investigated the effects of chance errors in the obtained correlation coefficients on subsequent factor analysis. A hypothetical factor matrix was constructed with four traits, twenty tests, and simple structure. "True" correlations were calculated from this matrix, and the "obtained" correlations, subject to chance errors, were derived by adding to the "true" coefficients errors varying from $-3\sigma_r$ to $+3\sigma_r$, distributed at random among the one hundred ninety coefficients. The standard errors of the correlation coefficients were calculated on the assumption of a population of one hundred. The "obtained" correlation matrix was factorized, and the centroid axes rotated to give a good fit to the true factor matrix. When the obtained factor loadings were compared with the true loadings, their standard error of estimate, *i.e.*, the root mean square deviation of the loadings, was .064, or somewhat less than the standard error of the average correlation coefficient, which was .084. There is no indication from this hypothetical situation as to whether the errors of the loadings would be larger for the less important factors, because the original factors were made about equally important. But the results do support our expectation that the factor loadings should vary less by chance than do the correlation coefficients from which they are calculated.

Mosier also considers the merits of several criteria for determining when the analysis has gone far enough. His best criteria were satisfied after four factors had been extracted. He demonstrates on this hypothetical problem that the errors of the factor loadings are much larger if the analysis is stopped after three factors have been extracted, but that after the necessary four factors are extracted the extraction of more factors does not significantly change the errors of the loadings.

2. STABILITY OF PRIMARY FACTOR LOADINGS

According to Thurstone's theory of analysis into primary factors, a given test must have the same primary factor structure regardless of the test battery with which it is associated (assuming, of course, that experience in the other tests does not alter the nature of the given

test). But factor analysis will only reveal, in each case, that part of the structure which is related to other tests in the battery. Thus, if performance in a certain test is determined by four factors, and other tests in the battery involve only two of these factors, only those two will be revealed. Consequently, the greater the variety of related tests included, the more likely is it that the analysis will reveal the complete factor structure of any test. This makes no assumption that the "primary" common factors represent unitary determiners, but merely that they represent foci of determination, or clusters of determiners that tend to occur together. It is important to realize that the "primary" factor structure is generally obscured in a centroid analysis. The first centroid represents an accumulation of the important determiners common to sets of measures; consequently, it reflects any change in the composition of the test battery. Thus, when we ask whether a test will have the same factor structure when appearing in different test batteries, we should not expect a crucial answer in terms of centroid factors. Yet, unfortunately, the frequently quoted investigation of this problem by Smart[16] was limited to incomplete centroid analysis.

Smart surmised that with a limited number of tests sampling behavior it would be improbable that the same factors would be found by analysis when the number of tests was altered. To test the proposition he used the intercorrelations of CA and ten measures of verbal and non-verbal ability, administered to a population of five-and-one-half-year-old children. Using various groupings of from four to eleven variables he made separate centroid analyses of each grouping and reported the first two centroid factors in each analysis. As more variables were added the factor loadings of given measures varied enormously, showing that the centroid factors depended upon the composition of the group of measures. (This is to be expected, of course, where measures of a different nature are being thrown together.) Smart concludes that one can not name centroid axes and depend on them to "stay put." In this he is quite right, but others may have implied that the same argument holds for all results of factor analysis. Dunlap[3] in his recent review of advances in statistical treatment cites Smart's study as the only evidence of constancy of factor loadings and leaves open the question as to whether the same inconstancy applies to rotated "primary" factors.

With this question in mind, Smart's analyses were repeated. His test measures will be referred to by the following abbreviations:

V—Minnesota Pre-School Scale, Verbal.
M—Minnesota Pre-School Scale, Non-verbal.
A—Arthur Point Scale of Performance.
S—Stanford Binet

McCarthy Language Survey:

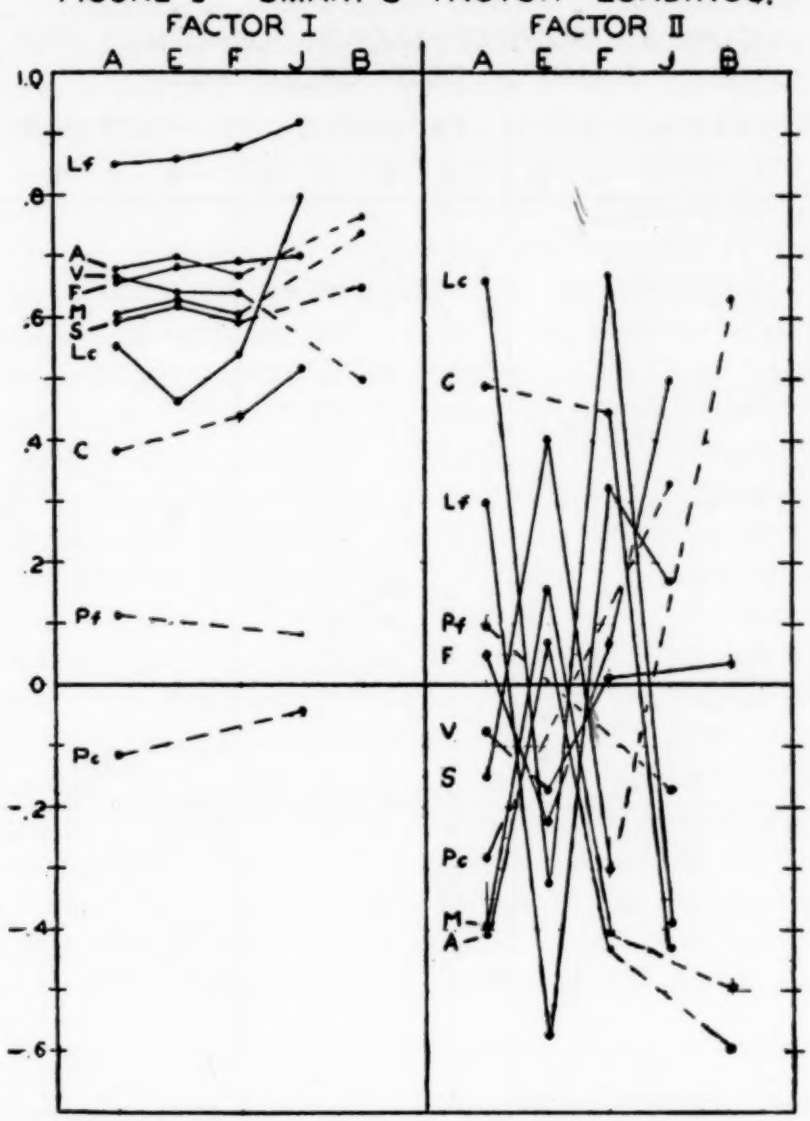
C—Average number of words in fifty controlled responses.
F—Average number of words in fifty free play responses.
Lc—Average number of words in five longest controlled responses.
Lf—Average number of words in five longest free responses.
Pc—Percentage of pronouns in controlled words.
Pf—Percentage of pronouns in free words.

An unfortunate feature of the investigation was that the tests were not all adiministered at the same age. The McCarthy Survey was administered at sixty-six months for all subjects, but the Minnesota and Arthur were given at from sixty-four to sixty-seven months, and the Stanford at from fifty-five to ninety-one months. Thus, chronological age was correlated with four measures and was recorded as correlating zero with the others, although mental maturity, which must have been largely represented by CA, was probably correlated with all of the measures of ability. This would undoubtedly distort factor analyses. The distortion can not be avoided, but in the repeated analyses reported below an attempt was made to minimize the cumulative effect of the distortion by partialling out CA, as a first factor with the following test loadings: V, .001; M, .145; A, .265; S, .700; and zero for all the other measures. Since the Stanford was given over such a wide age range, this partialling out of CA must have removed a considerable part of its mental-age factor, but the other tests were not much affected.

Figure 1 shows the first and second factor loadings of the various tests from Smart's analyses of groupings *A*, *F*, *E*, *J*, and *B*, involving respectively ten, eight, seven, six, and four measures (in addition to CA). The values are taken from Smart's Plates I to V.[16] The lines follow the factor loadings of a given measure through successive test groupings, a horizontal line indicating constancy. As Smart observed, the loadings of the second factor are extremely variable. The first factor remains fairly constant, but it does not allow a reliable distinction between measures *A*, *V*, *F*, *M*, *S*, and *Lc*. Admittedly the centroid analysis of these mental measures is confusing. When these factor loadings were plotted, it appeared that, by rotation of

axes, similar factors could be recognized in some cases where Smart had observed drastic changes in centroid loadings. There was no assurance, however, that two factors would in each case account for most of the common variance. Consequently, for these few groupings,

FIGURE I SMART'S FACTOR LOADINGS.

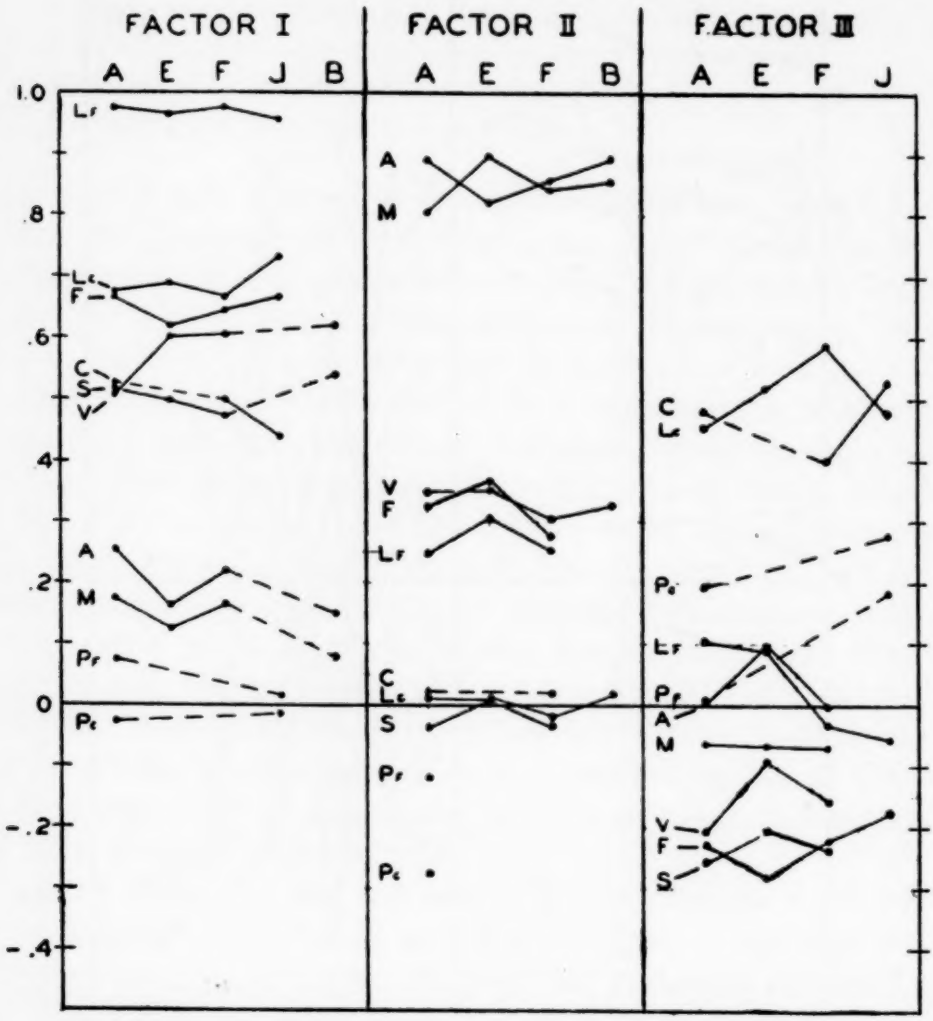


which Smart reports as showing extreme factor changes, the analyses were repeated, using a modified form of analysis[20] whereby orthogonal rotation is accomplished in the course of the analysis.

Figure 2 shows an analysis of the same test groupings in terms of three rotated axes, which account for nearly all of the common variance. Factor III does not appear in group *B*, and Factor II is not recognizable in group *J*, but the groups *A*, *E*, and *F* involve all three

factors. The first notable feature is that the same factors are recognizable throughout the various groupings of tests. A second important feature is that any given test maintains a fairly constant pattern of factor loadings regardless of what other tests it is grouped with. This is all as it should be, where simple "primary" factors are con-

FIGURE 2. ROTATED-FACTOR LOADINGS FOR SMART'S TEST GROUPINGS.



sidered. What variations do appear in the factor loadings may well result from the unreliability of the original correlations and from the small number of tests involved, (which tends to magnify errors from estimates of communalities).

It will be noticed that the new analyses more adequately distinguish between the various tests in that they now allow one to distinguish consistently the factor patterns of tests *A* and *M* from those

of *V*, *F*, and *S*, and of *Lc*. Whereas it would have been unsafe to name the centroid axes, it now seems fairly justifiable to interpret the rotated Factor I as related to verbal intelligence, Factor II as related to non-verbal intelligence (performance), and Factor III as involving performance in controlled- as opposed to free-situations. The interpretation is gratuitious, but it agrees plausibly with the ideas of many clinicians. We see, then, that the variability of centroid loadings does not necessarily apply to rotated "primary" factor loadings.

Although there is a tendency of convergence toward agreement in defining abilities, the field of personality characteristics has revealed much less consistent structuring. The slow progress may result in part from the tendency to perpetuate inadequacies of exploratory studies by continuing to use self-evident questionnaires. Another common fault has been the use of some of the same items in questionnaires purporting to measure neuroticism, adjustment, ascendance, and introversion. Factor studies have begun to separate out the various kinds of items involved, but there is not yet agreement upon the fundamental clusters of items, nor upon the interpretation of what characteristics they represent. In one case, however, there is close agreement between the Guilfords' "*S*"-factor (Seclusiveness),[5] from analysis of traditional *I-E* items, and Mosier's "*S*"-factor (Social Introversion),[10] from analysis of items of the Thurstone Neurotic Inventory. The Guilfords purposely included some "*S*"-items in a recent study of other aspects of personality, and they report[6] that the few items common to their two studies, and to Mosier's, emerged with similar factor loadings on an "*S*"-factor despite the large differences in the test batteries of items. If this indication is supported by further evidence it will encourage factorists to believe that they are approaching dependable focal aspects of personality, but this need not imply that the aspects remain constant in any given individual.

3. PATTERN VARIABILITY AMONG POPULATION SAMPLES

There is always the danger that the pattern of factor loadings for a test battery is peculiar only to the given group of subjects. If other apparently similar populations are tested, will the same factor pattern be found? Unfortunately, no published studies have been specifically aimed at this problem, but several give relevant evidence. In a study of rational learning by Roslow,[12] two similar classroom populations were used but the batteries of tests were only partly similar. Four of the learning problems were the same, but the fifth was a four-unit

problem for one group and a twelve-unit problem for the second group. The tests of intelligence and scholarship were different except for the Army Alpha and the Mathematics Content Tests. Roslow reported an analysis into three centroid axes for each group, and was perplexed by the dissimilarity of Factors II and III in the two groups. The chaos is structured a bit more clearly when the factor matrices are multiplied by the orthogonal rotation matrices in Table I. The resulting rotated factor matrices are shown in Table II. The first factor, for both groups, obviously pertains to test-intelligence, and is only correlated

TABLE I.—ROTATION MATRICES FOR ROSLOW'S FACTORS

	Group I Roslow's Factors				Group II Roslow's Factors		
	I	II	III		I	II	III
I′	.823	.541	.155	I′	.830	.541	−.132
II′	−.480	.816	−.324	II′	.540	−.712	.448
III′	−.300	.200	.933	III′	.150	−.440	−.885

TABLE II.—FACTOR LOADINGS OF ROSLOW'S TESTS
From Orthogonal Rotation of Centroid Axes

Factor	Group I			Group II		
	I	II	III	I	II	III
Learning problem						
A	.29	.32	−.09	.35	.46	−.02
B	.30	.79	.29	.22	.63	.32
C	.35	.61	−.08	.22	.70	−.06
D	.14	.81	−.19	.23	.66	.31
E	.18	.44	−.15			
E′	.....	.....	.....	.26	.81	.01
Scholarship						
One year	.....	.....	.....	.42	−.19	−.05
Two years	.59	−.03	−.03			
Army Alpha	.77	.17	.44			
Otis Self-administering	.51	.20	.38	.73	.11	−.25
Henmon Nelson	.....	.....	.....	.71	−.04	.08
Mathematics content	.64	.05	−.51	.76	.06	−.42
English content	.....	.....	.....	.42	−.06	.03
Thurstone Personality Schedule	−.01	.00	.52			

around .2 or .3 with the learning tests. Factor II for both groups correlates near zero with all measures except the learning problems. The justification of considering it as a special ability in rational learning tests seems strengthened by the fact that it correlates .44 with the four-unit problem (*E*) and .81 with the twelve-unit problem (*E'*). The third factor, however, seems to represent rather different residues in the two groups and can not be identified beyond the fact that it correlates negatively with mathematics. It seems probable that more than three factors are involved in these measures, but since those beyond the third were neglected the analysis may not have been adequate to allow clarification of factor three.

Another study of abilities, by Schiller,[14] serves our purpose better in that the same battery of tests was administered to a group of third- and fourth-grade children, and the results analyzed separately for the one hundred eighty-nine boys and for the two hundred six girls. These two groups may be considered comparable except for whatever influence sex may have on the organization of test abilities. Schiller's tetrad analysis revealed a strong general factor and minor verbal, numerical, and spatial factors, although the verbal and numerical factors were not clearly differentiated. More recently Garrett[4] repeated the analysis of the boys' data, using a centroid analysis. The fourth factor was discarded as unimportant and the first three factors were rotated into an oblique simple structure ([4], p. 270). The "verbal" and "numerical" factors correlated .825 with one another, but the "spatial" factor correlated less than .3 with the other two. These results, as Garrett says, show more clearly than did the tetrad analysis that the three primary abilities are not independent. Our interest now is to discover whether the relationship is similar in the girls' test data.

Comparison in terms of oblique axes might not be convincing, for Garrett's oblique axis loadings do not account for all of the common variance of the tests. The general factors common to all the tests are probably somewhat neglected by the oblique rotation procedure. Consequently, for safety, we shall compare orthogonal factor structures. For the boys, Garrett's four centroid axes ([4], p. 266) were rotated orthogonally into as simple a structure as seemed obtainable with independent dimensions. The girls' test correlations (from Schiller's monograph) were factorized into four centroid axes which were then rotated into a more simple orthogonal structure. The resulting factor loadings from the two analyses are compared in

Table III, which can be read as follows: The Number Series Test has a loading of .57 by the boys' Factor I, and has a loading of .63 by the

TABLE III.—FACTOR LOADINGS OF SCHILLER'S TESTS
From Orthogonal Rotation of Centroid Axes

Test	Factor I		Factor II		Factor III		Factor IV	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Number Series	.57	.63	.30	.25	.53	.37	.00	.07
Arithmetic Reasoning	.46	.63	.62	.44	.46	.35	.02	.07
Computation	.49	.42	.47	.46	.34	.48	−.21	−.07
Vocabulary	.39	.38	.77	.74	.11	.16	.06	.08
Analogies	.49	.56	.64	.56	.26	.15	.05	−.02
Sentence Completion	.47	.50	.73	.75	.25	.12	.07	.05
Reading	.38	.41	.80	.74	.00	.07	−.12	−.27
Otis	.75	.75	.13	.28	.04	−.20	−.07	−.05
Army Beta	.79	.69	.01	.09	.05	.26	−.18	−.10
International	.71	.72	.22	.31	.10	.07	−.20	−.23
Goodenough	.56	.44	.06	.23	.00	−.06	.22	.19
Performance	.57	.48	−.05	.16	.01	.08	.17	.29

girls' Factor I; the same test has a loading of .30 by the boys' Factor II, and of .25 by the girls' Factor II, etc.

There is a striking similarity in the factor patterns for the two groups. Consequently certain dissimilarities deserve notice as possible indicators of sex differences. Factor I obviously relates to intelligence as measured by non-verbal tests. This is probably not a simple, unitary ability but rather a cluster of the abilities sampled by the Otis, the Army Beta, and the International Intelligence Test. All of the tests involve some of these abilities and so have appreciable loadings by Factor I. But over and above this general cluster there are certain independent abilities. Factor II seems to represent ability in using a verbal medium, with heaviest weighting in reading, vocabulary, and sentence completion. But success in the numerical tests is also related to this factor, so that the remaining "numerical" Factor III is not very important. Thus we find, as did Schiller and Garrett, that there is not a clear differentiation between ability in the verbal and in the numerical tests. The most significant sex difference in factor loadings is that Arithmetic Reasoning appears more closely related to the non-verbal abilities in the girls, and more closely related to verbal facility in the boys, which may simply reflect the different social pressure on boys and girls, where arithmetic is concerned. Factor IV is unimportant, but seems to differentiate manipulative performance from other non-verbal abilities.

No attempt is being made to defend this factor structure as the most enlightening analysis of the test abilities. In fact, it seems unjustifiable to expect a psychologically significant analysis to result from the factorization of such complex measures as total intelligence test scores. The analysis would doubtless be clearer if one started with simpler subtest scores and investigated the extent to which these more specialized abilities remained separate or clustered together. The verbal and number tests come close to being simple measures, but their advantage is obscurred by correlating them with complex intelligence test scores. Quite apart from factor interpretations, however, the purpose of the above analyses was to demonstrate factor pattern constancy. No matter what the "unitary factor structure" of the tests may be, it can also be represented by the simplified orthogonal factor pattern. And it has been illustrated above that when similar groups are tested with the same test battery, factorization of the intertest correlations reveals fairly similar factor loading patterns.

Another illustration of this point will appear below in the discussion of the study by Rundquist and Sletto.

4. TIME VARIABILITY

The foregoing evidence, although scanty and inconclusive, certainly suggests that when a test battery is administered to similar groups, analysis will reveal quite similar patterns of factor loadings. Moreover, even when tests appear in different batteries they can often be shown to maintain the same primary factor structure. But may one expect the factor pattern for a given set of tests to remain constant over a period of time? The answer should be obvious if one remembers that the factor pattern is merely a clearer statistical representation of the intercorrelations of the tests. If the correlations do not change significantly over a period of time, then the factor structure of the tests must have remained the same. But, on the other hand, influences which change the correlations will change the factor pattern for the set of tests. This should not be considered an objection to factor analysis! Psychologists all know that the nature of a task may change with practice or as a result of some other influence upon the individual performing the task. For example, a performance which originally depends upon rapid adaptation to unfamiliar relationships may later depend mainly on endurance and steady attention. The point to be made here is simply that wherever there are changing relationships, factor analysis can be utilized as a means of more clearly showing the change and, perhaps, more logically describing the nature of the change, and this can be done whether one thinks of the factors as related to "real unitary traits" or merely as useful psychological constructs.

A study by Anastasi[1] was expressly designed to demonstrate that a practice period could change the interrelationships among a set of test scores. Analysis by the method of principal components was used to demonstrate the shift in factor pattern resulting from the practice experience. Unfortunately, Anastasi's use of the phrase "mental organization" had implications which she did not intend, and this, combined with the fact that she included only five tests in the study, brought forth criticisms from Thurstone[19] in a very lucid presentation of his views on how the factor methods should be used. Thurstone's criticisms, and Anastasi's reply to them,[2] serve to contrast the points of view of those who use factor analysis merely as an immediate description of a set of correlations, and those who seek to extend the

description back to more stable, universally recognizable variables. Either use is justifiable provided the user recognizes his orientation and follows methodological requirements. One of these is to use a sufficient number of tests to determine adequately the factor structure which one intends to offer as a representation of the test interrelations. Furthermore, if a comparison of factor patterns is the point at issue, one should offer some convincing argument that patterns are actually as similar as could be expected under the circumstances. Or, if a difference is at issue, it is not enough to show that the factor patterns look different, but it should also be demonstrated that the patterns can not be converted into similar patterns by rotation. We have seen above that centroid factor patterns may appear very different, yet when rotated into simple structure may be shown to represent fairly similar factor patterns. Consequently, the more convincing argument is to rotate centroid patterns into simpler patterns, if that is possible, and then to show that the simple structures are alike or different. The same argument holds when the analyses are made by the method of principal components, which have repeatedly been shown to yield results similar to those of centroid analysis. Anastasi's reply[2] does not make it apparent that she is aware of this real problem in comparing factor patterns. Actually, in view of the changes in correlations, there is little doubt that she has demonstrated a change in factor pattern, but the evidence is very scanty.

Another investigation which demonstrates practice effects is that of Woodrow,[21] who gave his fifty-six subjects thirty-nine days of practice on seven tests. The scores on these tests, before and after practice, and also several other "end-test" scores, were all correlated and the correlation matrix of thirty-three variables (including initial and final scores) was analyzed by the centroid method, followed by rotation to attain as simple a structure as possible. In this way the factor structure of initial and final tests could be directly compared in terms of the same common factors. The author states that the most important fact established is the marked change in factor loadings after practice, showing that a performance, after practice, may depend more upon certain abilities and less upon others than it did initially. It takes only a little imagination, then, to see how factor pattern changes may be very useful in demonstrating changes in the nature of a performance.* Thurstone would like to show that the underlying deter-

---

* A recent study by McNamara and Darley[9] has demonstrated that the

miners or abilities have recognizable continuity in time and in different situations, so that a given performance may depend now primarily on one of the abilities, and later on another of them. But other investigators may prefer merely to show the change in nature of the performance without making assumptions as to the constancy of underlying abilities. It is unfortunate that there is not a more widespread understanding of these diverse uses of factorial descriptions, for no doubt many investigators are deterred from using factor methods by the fear that it will commit them to a belief in rigid traits or faculties.

5. PATTERN VARIATIONS

Finally, one asks whether the factor pattern for a set of tests can be expected to remain constant from one group of subjects to another. Here again it is obvious that if there are group differences which result in altered correlations, then the factor patterns must also mirror the differences. Where many correlations are involved it is simpler and more convincing to compare factor patterns than to study correlation coefficients and partial correlations. Moreover, it is usually desired to show the nature of the differences, and simple-structure factors, if they can be given meaningful interpretations, permit the most efficient (or at least the simplest) description of the differences. One instance in which the same tests were administered to many different groups of subjects occurred in a study of Personality in the Depression, by Rundquist and Sletto,[13] in which they reported the intercorrelations of six attitude scales for each of the groups. Smart, in his aforementioned article,[16] presents centroid analyses for each of the groups and interprets the great variations in these factor patterns as demonstrating the unreliability and uselessness of factors discovered in different populations. As has been shown above, the argument in terms of centroid factors could not be very convincing, and as it served to obscure the useful information which factorization could have yielded, it may be worth while to reconsider the case in more detail, even though the number of variables is too small for safety.

The Rundquist and Sletto attitude scale is divided into six subscales—Morale (*M*), Inferiority Feelings (*I*), Family Adjustment (*F*), Respect for Law (*L*), Economic Conservatism (*C*), and Regard for Education (*E*). The scales were administered to a group of five

relations between attitudes and adjustment change somewhat in retest situations. With the help of factor analysis the nature of the change is sought and briefly described.

hundred standard men (SM) typical of the local population, and to groups of one hundred each of high-school men (HSM), college men in Sociology I (Soc. I M), control unemployed men (CUM), and men on public relief during the depression (DPR). The women's groups were similar except that there was no public relief group. In the present study to check upon Smart's implications it was decided to start with the correlations as presented by Rundquist and Sletto.[13] For each group the correlations were analyzed by a modified centroid method such as that described by Woodrow and Wilson.[20] The factors extracted were all orthogonal (independent), but by this method the orthogonal rotations were made in the course of the analysis rather than subsequently, and the express aim was to discover whether similar factors could be extracted from the data of the different groups. The factor patterns for the five male groups and for three of the comparable female groups are shown in Table IV. Admittedly, analyses based on only six variables are not extremely accurate, but the indications from these analyses were so interesting as to lead the writer more carefully to peruse Rundquist and Sletto's monograph. There it was found that much more detailed and laborious investigations confirmed several of these conclusions suggested by the factorizations. Consider, then, what the factor patterns indicate.*

First, the complexity of relationships between attitudes is indicated by the number of factors necessary to account for the intercorrelations. For the high-school men and the standard men two factors account for the relationships, with negligible residuals, and the factor patterns are almost identical for the two groups. This implies that the attitudes on the six scales could normally be reduced to only two types of attitude (and apparently the same ones) for high-school and adult men.

* Since the completion of this analysis another study has been reported by McNamara and Darley[9] using these attitude scales in combination with other attitude and adjustment questionnaires. They present a rotated factor pattern of five factors, three of which are quite similar to those isolated in our analysis, for college men. Where there are discrepancies a hurried analysis shows that a bit of rotation can bring our factors closer to theirs, or *vice versa*. In view of the greater number of tests included in their study, their factors should allow a more accurate analysis of common relationships, but fortunately the differences are not great enough to alter the present interpretation of group differences in Rundquist and Sletto's study. McNamara and Darley, using both men and women subjects, were studying the change of organization of attitudes with the passing of time. They report several changes in factor pattern in the retest situation, and from the original and retest factor patterns they interpret briefly the nature of the changes in organization of attitudes, without making assumptions of constant traits.

Assuming (point 2, above) that the same group of tests administered to similar groups of subjects will reveal similar factor patterns, the corollary here is that when two groups show similar factor patterns, the groups must be alike with respect to the organization of functions revealed by those tests. The organization of attitudes becomes increasingly more complex in the other groups, and additional common factors and specifics are necessary to account for the relationships. Admittedly the six variables can not give an adequate determination of more than two factors, but they can reveal which of the relationships require additional determining factors. Moreover, one can probably obtain a fairly good indication of the importance of additional factors by applying the method here used of guessing at the nature of the factors and rotating during the process of extraction.

From several kinds of item analysis Rundquist and Sletto conclude that women's attitudes are not as consistently organized around a central core as are men's. The factor patterns show this at a glance, in that both the high-school and standard women require much more complex factor patterns than do the corresponding men's groups. Because of this initial complexity of women's attitudes the changes of organization in the unemployed women do not seem nearly as extreme as the changes from simple organization in the standard men to complex and specific relationships of attitudes in the unemployed and public-relief men. This agrees with the conclusion of Rundquist and Sletto that women's attitudes are not affected as much as men's by unemployment.

In addition to the complexity of organization, the nature of the changes in attitudinal factors is indicated by the factor patterns. For such interpretations, however, it helps to examine the items included in each scale, for the division of scales is sociological rather than psychological. The Morale scale includes attitudes both of discouragement and of cynicism regarding society. The Inferiority scale includes a few items of discouragement. The Family scale includes attitudes of helpless submission and of resentment of control; the Law scale involves both distrust of others and resentment of control; and the Education scale is heavily weighted with cynicism and disillusionment. One might expect, then, to find a factor common to the *M*, *F*, *L*, and *E* scales due to their related expressions of discouragement and disillusionment. And such must be the nature of the men's Factor I, which is independent of conservatism-radicalism, and is related only slightly to inferiority feelings. Incidentally, this factor

corresponds closely to the nature of the sixteen most discriminative items selected by Sletto by the method of internal consistency.

TABLE IV.—FACTOR PATTERNS FOR RUNDQUIST AND SLETTO ATTITUDE SCALES, FOR VARIOUS GROUPS

Scale	HS men Factor		Standard men Factor		Sociology I men Factor		
	I	II	I	II	I	II	III
M	.60	.57	.65	.52	.63	.52	.33
I	.26	.59	.26	.56	.32	.46	.20
F	.65	.02	.58	.01	.58	.06	.01
L	.63	.18	.62	.23	.48	.40	−.03
C	.01	.34	.09	.38	.13	.42	.03
E	.63	.12	.63	.09	.50	.02	.48

Scale	*C.* Unemployed men Factor					DPR men Factor			
	I	II	III	*Sp*	*Sp*	I	II	III	*Sp*
M	.63	.50	.44	.35	...	.60	.58	.48	
I	.30	.42	−.03	.40	...	.64	.48	.01	
F	.58	.00	−.01	...	...	.69	.00	−.04	
L	.56	.02	.40	...	.42	.54	.02	.54	.65
C	.19	.39	.05	...	.42	.14	.51	.09	.65
E	.47	.03	.40	...	...	.39	.02	.63	

Scale	H-S women Factor				Standard women Factor				*C.* Unemployed women Factor			
	I	II	III	*Sp*	I	II	III	*Sp*	I	II	III	IV
M	.62	.60	.40	...	.58	.54	.36	...	.59	.42	.50	−.03
I	.60	.40	−.03	...	.44	.40	.00	...	.60	.50	.00	.02
F	.71	.00	.02	...	.66	.00	.00	...	.66	.00	.00	.52
L	.58	.01	.40	.38	.58	.04	.36	.40	.40	−.02	.44	.50
C	.21	.35	.24	.38	.20	.39	−.01	.40	.13	.29	.38	.01
E	.48	−.01	.49	...	.44	.05	.44	...	.42	.00	.60	.30

Inspection of these items ([13], pp. 220–221) reveals that they all involve discouragement or disillusionment regarding people or institutions.

It can not be argued that this is a unitary factor (the proof is insufficient) but merely that it represents a cluster of attitudes which hold together in the same way in the first four men's groups. Actually the "disillusioned" attitudes toward Law and Education become less closely related to this factor for the Sociology I men and the unemployed men, but the real shift in Factor I occurs in the DPR group, in which attitudes of Inferiority become more closely related to disillusionment and family adjustment.

Factor I for the DPR men has the same pattern as the women's Factor I, which leads us to the sex comparisons. Rundquist and Sletto conclude ([13], pp. 329–334) that women's attitudes are organized around a completely different core (Family and Law) than are men's, in that different items are most discriminative for the women. As they have shown, however, this difference may be partly due to specific wording of the items. Their list of ten most discriminative items for the women involve dissatisfaction with home (*F*), disillusionment (*L*, *C*), and poor adjustment with intimates (*F*, *I*). (The one *C*-item shows disillusionment rather than radicalism.) Consequently it seems sensible that the Morale scale, which involves all three of these elements, should be heavily loaded by our Factor I, representing these "personal adjustment" attitudes. The factor pattern suggests not that there is a different core of women's attitudes, but rather that the Inferiority attitudes are much more closely related to Family adjustment for the women. The distinction is especially clear in the high-school group. For the standard women, Inferiority becomes less related to the personal adjustment aspect of morale (the Factor I loading drops from .60 to .44), but for the unemployed women there is a regression, probably as a result of their being forced back to more dependence upon the family. It seems plausible, then, that a similar effect upon attitudes should be observed in the DPR men, who likewise tend to be thrown back into the home and become more dependent upon the family.

Factor II is recognizably similar in all groups, male and female, in that it involves attitudes common to Morale, Inferiority, and Conservatism. Inspection of the items suggests that this must represent feelings of inadequacy, or projection of the inadequacy to social institutions, and the desire for radical revision of the institutions. For the first three men's groups, this factor also appears in the Law scale, which includes some items regarding desire for freedom to better one's own welfare. For the other groups, however, Table IV shows that

the relation between legal and economic radicalism (*L* and *C*) becomes more specific and is independent of the Morale scale. But there is no assurance that this specific factor for the unemployed men is the same as that for the women. In fact, these interpretations are handicapped throughout by the small number of scales and by the fact that the scales are not homogeneous, *i.e.*, do not represent empirical clusters of attitudes.

Factor III begins to appear in the Soc. I Men as a separate variety of cynicism related to the value of education. For the unemployed and DPR men and for the women there is a consistently recognizable Factor III representing attitudes toward the social institutions of Law and Education, and related, of course, to the *M* scale. This factor may represent disillusion, as in Factor I, but it shows a distinction between one's personal adjustment and one's attitudes toward more distant social institutions. The distinction seems to become more marked as the individual is subjected to more diverse social pressures. The unemployed women even have a fourth factor suggesting that some of their attitudes toward Family, Law, and Education have become divorced from their general morale. Such increased differentiation of attitudes does not necessarily indicate disorganization of personality, but may be merely a result of discriminative learning.

SOME ADVANTAGES OF FACTOR METHODS

This lengthy discussion of a single study has been introduced to exemplify the various ways in which factor patterns may be useful in interpreting changing relationships. To notice only the inconstancy of centroids and to overlook all the interesting implications of the variations is to throw away useful information. Factorization of correlations can not take the place of more detailed analysis of raw data, but it can reveal many possibly fruitful relationships more efficiently than certain other methods. For example, the detailed and laborious analysis of the attitude scales by partial correlation[13] left a blurred picture of the nature of the relationships between scales. The factor patterns suggest the nature of these relationships more clearly and with much less labor, and, since neither method of analysis is conclusive, the less tedious one would seem preferable. Still another point—the factor pattern reveals what Rundquist and Sletto suspected, that the six scales are not independent or unitary clusters of attitudes, but that the various clusters of attitudes cut across the scales. Sletto ([15], p. 71) concludes that many items could belong to several scales, and

that the method of internal consistency will not yield homogeneous scales where the scales include two or three types of item, for items of each type could correlate equally with the composite and thus be retained. In view of this deficiency in the criterion of internal consistency, which has been made even more explicit by the detailed study of Rundquist and Sletto, it would seem that an empirically more significant subdivision of attitudes could have been attained by analysis of inter-item correlations, and with perhaps little more work than was involved in calculating discrimination values for all items on each of the *a priori* scales. Our factor patterns suggest the probable nature of these empirical clusters of attitudes and also show that the clusters would differ in groups of different sex or economic status. Certainly, then, it should not be argued against such uses of factor analysis that they force one to accept a rigid and static concept of "traits," for they can equally well be used merely to reveal clusters of behavior items which operate together under certain sampling conditions.

SUGGESTIONS FOR A MORE FLEXIBLE USE OF FACTOR METHODS

The very enlightening new treatise by Thomson[17] discusses many possible applications of factorial methods, yet they generally seem to assume permanent factors. It is probably this assumption that has led investigators to doubt the utility of factorial methods in studying personality, which appears more changeable than abilities.[17,19] Yet factor patterns, as descriptive constructs, might well have utility in that field. With psychology becoming increasingly "dynamic" it would appear unwise to handicap investigations by the use of static techniques. Certainly to many psychologists it has seemed that factor analysis aimed at the discovery of unitary (and probably permanent) abilities or traits is in danger of being such a static technique, and consequently to be avoided. Strictly speaking it probably is a static method, like so many scientific methods, but it may nevertheless be of some use even in a dynamic psychology. One aim of the present discussion has been to suggest that constancy is not the only property to be admired in factor patterns. Variations may be equally significant and even more useful in describing or interpreting the flux of psychological processes. Before condemning the method it would be well to investigate further its utility for purposes such as the following:

(*a*) To discover empirical clusters of behavioral elements (or item responses). For this purpose other methods of cluster analysis, such as correlation profile analysis, may prove simpler or more comprehensi-

ble to some investigators, but the mathematical rationale of orthodox factorization is probably easier to defend.

(*b*) To isolate focal aspects of personality or behavior. If the empirical response clusters represent the confluence of several determining influences, and have boundaries determined fortuitously by the sampling of behavior, then factor analysis provides a means of investigating aspects which cut across several clusters (determining tendencies common to several clusters). Admittedly these focal aspects will be psychological constructs whose utility must be further verified.

(*c*) To describe differences in mental organization in different individuals or groups. Factor patterns are useful if the relationships can be described in terms of a limited number of aspects. These aspects could be different in each individual, or they could be common only to individuals of a certain type, but most factorists, to date, have assumed that the aspects should be the same for all individuals. Any of these views might lead to useful results without being universally correct.

(*d*) To reveal and describe changes of organization of mental functions resulting from social influences, learning, or changes in motivation.

(*e*) To test the empirical validity of various theories concerning organization of behavior. If *a priori* factor loadings can be derived from some theory, one can extract such factors from correlation matrices, testing the adequacy of the postulated factors or discovering what residual relationships remain to be explained. This may help in leading to fresh concepts, or to the revision of old concepts.

Factor methods, like correlation methods, are suited to preliminary investigations for discovering the most promising variables for further study. No one expects factor analysis to be a talisman capable of dissolving the many walls of present ignorance, but it may yet be shaped into a more useful tool for hacking at them. To the present writer it would appear that the utility of the methods may be increased by applying them more flexibly, by recognizing their results as descriptive constructs and, therefore, tentative; and by utilizing both constancy and variation of factor patterns in seeking to interpret functional relationships.

SUMMARY

Factor patterns are here considered as convenient statistical descriptions of the relationships between many variables. Ideally they

should provide a description which is more meaningful and simpler to comprehend than the observed correlation coefficients. The present discussion concerns the reliability of "primary" factors or of simple rotated factor patterns, and the possible sources of their variation. Most of what is said does not apply to centroid factors, which, in many circumstances, are neither dependable nor psychologically interpretable.

Logical considerations and the available empirical studies indicate that a factor loading will vary less with random sampling than will the correlation coefficients from which it is calculated. This assumes, of course, that the factor is fairly important, for the reliability of a factor loading depends upon the importance of the factor, as well as upon the magnitude of the observed correlations, the size of the population, the number of variables involved, and the errors in estimating communalities.

Evidence is offered to demonstrate that the factor loadings of a test remain the same when the test appears in different batteries of tests involving some or all of the factors.

Several analyses show that where similar populations are tested similar rotated factor patterns can be revealed. When the organization of abilities or traits changes, however, the factor patterns also reflect these changes. It is illustrated that the factor pattern variations are useful in describing the nature of the changing relationships.

It is suggested that more flexible applications of factorial methods would greatly increase their utility in psychological studies, especially where it is possible that the focal aspects (factors) being measured are not the same in all individuals, or do not remain constant in the same individual.

REFERENCES

1. Anastasi, A.: "The influence of specific experience upon mental organization." *Genet. Psychol. Monogr.*, Vol. XVIII, No. 4, 1936, pp. 245–355.
2. Anastasi, A.: "Faculties versus factors: a reply to Professor Thurstone." *Psychol. Bull.*, Vol. XXXV, 1938, pp. 391–395.
3. Dunlap, J. W.: "Recent advances in statistical theory and applications." *Amer. Jour. Psychol.*, Vol. LI, 1938, pp. 558–571.
4. Garrett, H. E.: "Differentiable mental traits." *Psychol. Record*, Vol. II, 1938, pp. 259–298.
5. Guilford, J. P., and Guilford, R. B.: "Personality factors *S*, *E*, and *M*, and their measurement." *Jour. Psychol.*, Vol. II, 1936, pp. 109–127.
6. Guilford, J. P., and Guilford, R. B.: "Personality factors *D*, *R*, *T*, and *A*." *Jour. Abn. & Soc. Psychol.*, Vol. XXXIV, 1939, pp. 21–36.

7. Kelley, T. L.: *Essential Traits of Mental Life.* Harvard University Press, 1935.
8. Lorge, I., and Morrison, N.: "Reliability of factor loadings." *Science*, Vol. LXXXVII, 1938, pp. 491–492.
9. McNamara, W. J., and Darley, J. G.: "A factor analysis of test-retest performance on attitude and adjustment scales." *Jour. Educ. Psychol.*, Vol. XXIX, 1938, pp. 652–664.
10. Mosier, C. I.: "A factor analysis of certain neurotic symptoms." *Psychometrika*, Vol. II, 1937, pp. 263–286.
11. Mosier, C. I.: "Influence of chance errors on simple structure: An empirical investigation of the effects of chance error and estimated communalities on simple structure in factorial analyses." *Psychometrika*, Vol. IV, 1939, pp. 33–44.
12. Roslow, S.: "A statistical analysis of rational learning problems." *Jour. Genet Psychol.*, Vol. XLVIII, 1936, pp. 441–467.
13. Rundquist, E. A., and Sletto, R. F.: *Personality in the Depression: A Study in Attitude Measurement.* Institute of Child Welfare Monograph 12, Univ. Minn. Press, Minneapolis, 1936.
14. Schiller, B.: "Verbal, numerical, and spatial abilities of young children." *Arch. Psychol.*, No. 161, 1934, 69 pp.
15. Sletto, R. F.: *Construction of personality scales by the criterion of internal consistency.* Sociological Press, Minneapolis, 1937.
16. Smart, R. C.: "The variation in patterns of factor loadings." *Jour. Educ. Psychol.*, Vol. XXVIII, 1937, pp. 55–64.
17. Thomson, G. H.: *The Factorial Analysis of Human Behavior.* Houghton Mifflin Co., 1939.
18. Thurstone, L. L.: "Current misuse of the factorial methods." *Psychometrika*, Vol. II, 1937, pp. 73–76.
19. Thurstone, L. L.: "Shifty and mathematical components." *Psychol. Bull.*, Vol. XXXV, 1938, pp. 223–236.
20. Woodrow, H., and Wilson, L. A.: "A simple procedure for approximate factor analysis." *Psychometrika*, Vol. I, 1936, pp. 245–258.
21. Woodrow, H.: "The relation between abilities and improvement with practice." *Jour. Educ. Psychol.*, Vol. XXIX, 1938, pp. 215–230.

COLLEGE DEGREES AND ELEMENTARY-SCHOOL INTELLIGENCE QUOTIENTS

F. J. ADAMS

University of Texas

To what extent do elementary-school children who later obtain college degrees differ in intelligence from their classmates? Are the college freshmen who later complete the requirements for a degree a good cross-section of the freshman class in terms of intelligence, or are they a select group? Are the intelligence ratings of students obtaining the different baccalaureate degrees about equal? The present discussion seeks to contribute added information with respect to these and similar questions.

In the Spring of 1926 the writer administered the National Intelligence Test, Scale *A*, Form I, to fifteen hundred five pupils located from the low fourth through the high sixth grade in the ten white non-Spanish-American public elementary schools of a Texas college town. Since about five years ago, an attempt has been made to trace these individuals and to study their academic progress.[1]

Without entering into a discussion concerning the constancy of the IQ, and avoiding the necessity of establishing the accuracy of an IQ obtained from a single group intelligence test, the writer wishes to examine such relationships as may exist between selected items of information obtained from college records and the data resultant from such a minimum testing program.[2]

GENERAL COLLEGE RECORD

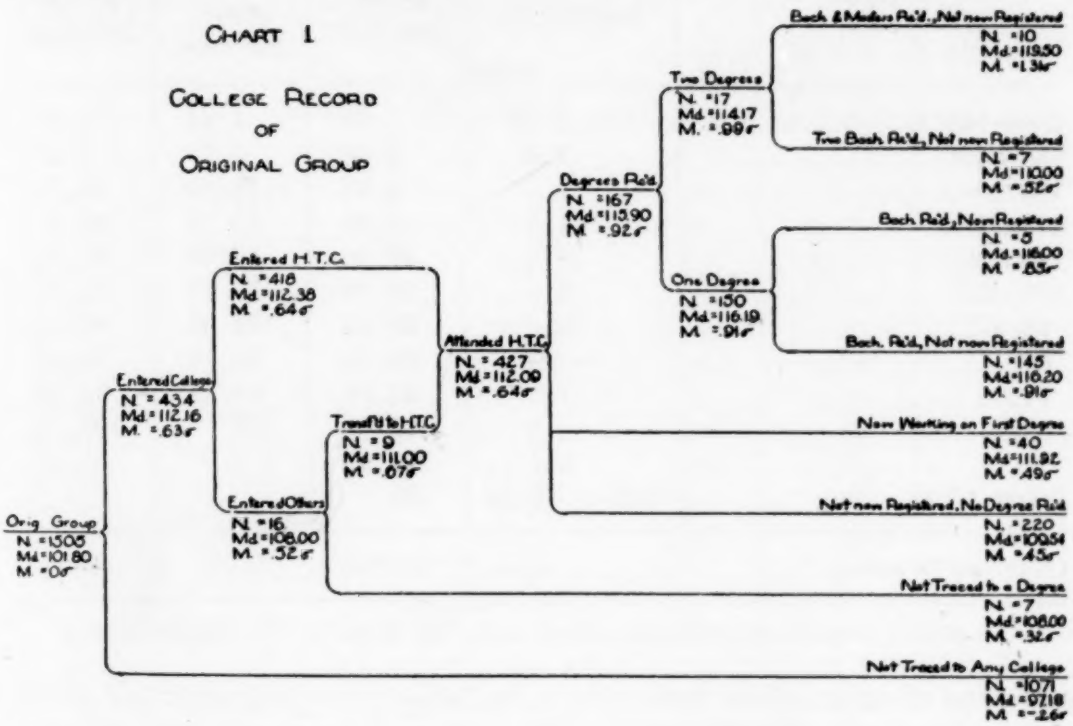
Of the original elementary-school group of fifteen hundred five pupils, the four hundred thirty-four who were traced to college obtained a median IQ of 112.16 in 1926, while for their classmates who were not traced to college a median IQ of 97.18 was found. This difference is statistically reliable ($D/\sigma_{diff.}$).

Chart 1 presents in graphic form the history of the original group of elementary-school pupils, as they have been traced into college and have progressed in their academic work. For each of the groups and subdivisions are stated the number of cases represented, the median

[1] Adams, F. J.: "Predicting High-school and College Records from Elementary-school Test Data." *Journal of Educational Psychology*, Vol. XXIX, 1938, pp. 56–66.

[2] The clerical and statistical services of J. D. Ford and Jack Presley, made available through the National Youth Administration, are gratefully acknowledged.

1926 IQ (hereafter referred to as the IQ of the students), and the mean IQ of the members of each group expressed in terms of its sigma equivalent in the original group's distribution of IQ's. The majority of the students traced to college entered their home town college (H. T.C.), while nine who began their college work elsewhere transferred to the institution in their home town. The remainder of this discussion will deal with the members of the H.T.C. group. Statistical reliability is found for the differences between the median IQ of the group composed of students who have already received degrees and



that of the degreeless group not at present attending college; also between the group with degrees and the group now registered for work towards their first degree. There are ninety-eight chances in one hundred that the median IQ of the students obtaining both a bachelor's and a master's degree differs from that of the students who obtained two bachelor's degrees.

SELECTIVITY

While Chart 1 indicates that the original elementary-school group, the group entering college, and the group obtaining one or more degrees are progressively smaller in size and progressively higher in median IQ, little can be judged therefrom as to the extent to which selection at the

different intelligence levels has taken place between these successive groups. Table I presents the distributions of these three main groups

TABLE I.—CUMULATIVE DISTRIBUTIONS, ORIGINAL, COLLEGE ENTRANCE, AND DEGREE GROUPS

Original group		Distributions in cumulative percentages		
IQ's	Sigma[1] equivalents	Original group, per cent	College entrance group, per cent	Degree group, per cent
Over 145.72	Over 2.5σ	.60	1.84	3.00
136.86	2.0	2.46	5.07	9.56
128.00	1.5	6.91	15.45	25.73
119.14	1.0	15.35	38.12	40.10
110.29	0.5	30.50	56.00	66.45
101.43	0	50.90[1]	79.72	89.80
92.57	−0.5	69.63	92.63	97.60
83.71	−1.0	84.18	98.16	98.80
74.85	−1.5	92.89	99.31	99.40
65.99	−2.0	97.41	100	100
57.14	−2.5	99.47		
Below 57.14	Below −2.5σ	100		
(Number of cases)		(1505)	(434)	(167)

[1] Standard deviations were computed from the mean of the original group.

in terms of cumulative percentages, by half-sigma step-intervals of the original group's distribution of IQ's. It will be noted that the students entering college and the members of this group obtaining degrees extend over a considerable portion of the range of intelligence quotients found within the elementary-school group from which they came. However, a comparison of the cumulative percentages tends to suggest that quite a little selection has taken place; about four-fifths of the group entering college came from the upper half of the elementary-school group, and about nine-tenths of the students obtaining degrees were in the upper half of the original group in the elementary grades.

Utilizing the same step-intervals within the original distribution of IQ's, Table II portrays these data in terms of the ratio between the relative frequency of students of the separate IQ levels in the college entrance group and the proportion of such students in the original

group; the comparison of the degree group and the original group; and the comparison of the degree group and the college freshman group. If both the group entering college and the students obtaining degrees were merely chance-selected members of the previous distri-

TABLE II.—COMPARISON OF ORIGINAL, COLLEGE ENTRANCE, AND DEGREE GROUPS, BY RELATIVE FREQUENCIES AT IQ LEVELS

Original group		Ratio of relative frequencies within groups		
IQ's	Sigma equivalents	College entrance/ original	Degree group/ original	Degree group/ college entrance group
Over 145.72	Over 2.5σ	3.23	5.00	1.66
136.86	2.0	1.73	3.55	2.04
128.00	1.5	2.33	3.63	1.56
119.14	1.0	1.50	1.70	1.14
110.29	0.5	1.84	1.74	.95
101.43	0	1.16	1.15	.98
92.57	−0.5	.69	.42	.60
83.71	−1.0	.38	.08	.21
74.85	−1.5	.13	.07	.52
65.99	−2.0	.15	.13	.87
57.14	−2.5			
Below 57.14	Below −2.5σ			

Read the table thus: When each of the three groups (original, college entrance, and degree) is proportionately reduced to the same number of students, the college entrance group contains 3.23 times as many students over 2.5 sigmas on the original IQ distribution as does the original group, the degree group contains five times as many students with IQ's over 2.5 sigma as does the original group, and the degree group contains 1.66 times as many students above the 2.5 sigma IQ level as does the college entrance group.

butions of intelligence quotients, then these ratios would uniformly approximate unity. Where an intelligence level contains a greater proportion of the later group than of the earlier group with which it is compared, the ratio exceeds unity. It may be of interest to note that IQ's less than 101.43 tend to be proportionately less frequent in the freshman and in the degree groups than in the original elementary-school group, while IQ's above 119.14 tend to be more frequent in the degree group than in the group entering college.

THE LOW IQ CASE

Of the fifteen hundred five members of the original group, one hundred seven obtained intelligence quotients below the −1.5 sigma position in the distribution—below 74.85 IQ. Three of these entered college, and one of the latter received a bachelor's degree. The case of this student may be of interest.

M. G. was fourteen years and three months of age when she was tested near the middle of the second semester of the sixth grade of elementary school. At that time, the test gave her a mental age of nine years and five months, and an intelligence quotient of sixty-six. She entered college in her home town when she was almost twenty years old, attended college the equivalent of ten semesters and obtained her bachelor's degree at the Summer commencement just before her twenty-fifth birthday. Her major subject was the foreign language usually associated with her family name. The college records credit her with fifteen semester hours of *A* grades (all in her major subject and six of these on freshman courses taken after her seventh semester of college attendance), forty-two semester hours of *B*, fifty-one of *C*, twenty of *D*, and thirty-one semester hours of work not passed. She undertook no work in mathematics or the science fields beyond the minimum required for her degree. Her record shows registrations for fifty-seven semester hours of freshman-level courses, of which forty-eight were passed; sixty-three sophomore course hours with fifty-one passed; thirty-five junior hours with twenty-nine passed; and four semester hours of registration in senior-level courses with none passed. After registering for one hundred fifty-nine semester hours of work, she was able to meet the grade-average requirement on one hundred twenty of these hours the degree demanded. To what extent her difficulty both with respect to the intelligence test and with reference to her college work may have been caused by a language handicap is not known, but some influence seems likely, as her father employed quite broken English when the writer recently checked with him as to the date of the student's birth.

IQ'S AND DEGREES OBTAINED

The one hundred sixty-seven students of the degree group received a total of one hundred eighty-four degrees. Table III permits a number of comparisons between divisions of the degree group, as it presents the median intelligence quotient and the sigma equivalent

TABLE III.—IQ's AND DEGREES OBTAINED

Degrees received	Number of cases	Median IQ	Sigma (original group) equivalent of mean IQ
Master's degree	10	119.5	1.31σ
All bachelor's degrees	174	115.0	.85
Bachelor of Arts	80	119.5	.93
Business administration degree	35	113.0	.86
Education degree	20	110.5	.63
Engineering degree	17	113.0	.68
Other bachelor degrees[1]	22	119.5	.91
Bachelor's degrees and honors			
Honors not available	44	112.0	.73
Honors available	130	115.5	.89
Highest and high honors	14	129.0	1.43
Honors	17	117.0	1.32
Honors not received	99	114.0	.84

[1] Degrees in architecture, geology, home economics, journalism, nursing, and pharmacy.

(within the original elementary-school group's distribution) of the mean IQ for students in the more frequent degree fields and with respect to honors being received with the bachelor's degree. While the frequencies of cases are too small for definite conclusions, a tendency seems to exist within this group of students for those obtaining a degree in education, engineering, or business administration to have been relatively less successful on the intelligence test—in fact, it may be recalled that the median IQ of the students entering the college here considered was 112.16, which does not differ greatly from the medians for these three degree groups.[1] If a larger number of cases had been basic to these data, it might be implied that students receiving bachelor's degrees with honors tend to possess, on the average, intelligence quotients above the majority of their classmates.

[1] These differences between the median IQ for Bachelors of Arts and students obtaining degrees in education, engineering, and business administration are not statistically reliable, there being, respectively, ninety-nine and eighty-nine hundreths, ninety-eight and four-tenths, and ninety-nine chances out of one hundred that the actual differences would exceed zero. When the reliability of the differences between the median IQ's for these three degree groups and the median IQ of the group entering college is computed, it is found that there are seventy-four and eight-tenths, sixty-three and eight-tenths, and sixty-four and six-tenths chances out of one hundred that the actual differences between the three degree groups, respectively, and the college entrance group would exceed zero.

IQ'S AND THE RATE OF SCHOLASTIC PROGRESS

A statistically reliable difference having been found between the median IQ's of the group obtaining a degree and of the degreeless group still attending college, it seemed of interest to compute the correlation coefficients between the intelligence quotients and various measures of the rate of scholastic progress for the one hundred sixty-seven students possessing degrees. If the obtaining of a bachelor's degree is viewed as representing the completion of the fifteenth year of study (a four-year degree curriculum following graduation from an eleven-year school system), the ratio of the *time required* by the student, since the elementary-school testing, to obtain the bachelor's degree to the *number of curriculum years* between his elementary-school location at the testing time and the completion of his degree work will furnish an index of his rate of progress for this portion of the student's educational history. Such ratios for the one hundred sixty-seven degree students and their IQ's correlate .19 $\pm$ .05. When a similar index of progress for the interval between the testing and high-school graduation is correlated with IQ, a coefficient of .05 $\pm$ .06 is obtained. The time intervals between graduation from high school and college entrance correlate .08 $\pm$ .06 with the IQ's of the students. The number of months, including the Summer-session period, between the admission of the students as college freshmen and the commencements at which their first or only bachelor's degrees were granted correlate $-.18 \pm .05$ with their intelligence quotients obtained during the elementary-school years. It would seem from these data that, within the group of students obtaining a baccalaureate degree, the intelligence quotients of the students, determined while they were in the later years of elementary school, are not very reliable as bases for predictions concerning the subsequent rate of scholastic progress of such students.

STUDENTS OF LESS THAN 100 IQ RECEIVING DEGREES

Four male and nine female members of the group receiving a baccalaureate degree obtained intelligence quotients of less than one hundred at the time of the elementary-school testing. In addition to the "low IQ case" already mentioned with the recorded IQ of 66, the IQ ratings of this lower group include 84, 91, 92, 94, 95, three members with 97, 98, and three persons with an IQ of 99. Their degrees were: B.A. with majors in English, Botany, Physics, Sociology,

and for two students modern foreign languages; B.S. in Education; B.S. in Physical Education for two students; B.S. in Nursing; Bachelor of Business Administration; B.S. in Civil Engineering; and B.S. in Mechanical Engineering.

When the thirteen lowest IQ individuals of the degree-receiving group are compared with the thirteen highest IQ members of this group, suggestions as to the accuracy of these lower IQ ratings are possible. The lower IQ group's median age at the time of college graduation was twenty-one years and nine months (range, 20-8 to 24-9), while the higher IQ group's median age was twenty years and one month (range, 19-1 to 22-3). In terms of the number of months between their admission as freshman students and their graduation, the lower IQ group's median time was fifty-seven months (range, 36 to 72), while the higher IQ group possessed a median of forty-five months (range, 39 to 69). No member of the lower IQ group either received honors at graduation, although seven of the thirteen individuals followed curricula with respect to which honors were possible, or continued college work after receiving their baccalaureate degree. Ten of the members of the higher IQ group followed curricula wherein honors were possible, eight of these receiving honors, including four highest honors, and five members of the higher IQ group have received master's degrees.

While it is not claimed that a single group test of intelligence administered during the elementary-school period will produce data from which accurate predictions can be made either respecting the extent or the rate of progress of achievement in college, the comparisons just cited seem to imply that the college achievement of lower IQ and of higher IQ students receiving college degrees tends to agree with their relative IQ standing.

CONCLUSIONS

Within the group of individuals with which this report is concerned, if the intelligence quotients derived from but one group intelligence test are assumed to be accurate evidence of the relative intelligence of the individuals, the following conclusions appear to be justified:

(1) Elementary-school students later entering college tend to be of higher IQ than their classmates who are not traced to college, and college freshmen who obtain bachelor's degrees average a higher IQ than the students who drop out of college before a degree is received.

(2) College students still working on their first degrees after their former elementary-school classmates have obtained degrees tend to be of lower IQ than their degree-possessing former classmates.

(3) There seems to be some possibility that students who choose as their second degree another bachelor's degree rather than a master's degree may be lower in IQ than the master's degree possessors.

(4) About four-fifths of the college freshmen and nearly nine-tenths of the students obtaining bachelor's degrees rank above the average IQ of their classmates in the later elementary-school grades.

(5) IQ's above 101 seem to be relatively more common among college freshmen and among students obtaining bachelor's degrees than among students in the higher elementary-school grades, and IQ's above 119 appear to be proportionately more frequent among college graduates than among college freshmen.

(6) The average IQ's of students obtaining the different bachelor's degrees would not seem to be identical; possibly the IQ's of the typical students in the fields of education, engineering, and business administration being lower than the average IQ of the Bachelors of Arts.

(7) Within the degree groups where honors can be granted, the data suggest that the students who obtain degrees with honors may possess higher IQ's, on the average, than their classmates who failed to receive honors with their degrees.

THE RELATIONSHIP BETWEEN BASAL PHYSIOLOGICAL FUNCTIONS AND INTELLIGENCE IN ADOLESCENTS*

NATHAN W. SHOCK AND HAROLD E. JONES

University of California

A marked deficiency of thyroid secretion, as in cretins, has long been known to be associated with mental and physical retardation.[6,10,15] Numerous speculations have also been made concerning the effects of hypersecretion, as in the statement by Crile:[4] "So high is the scholastic record among patients with hyperthyroidism and so many individuals of Phi Beta Kappa rank are to be found among them that although hyperthyroidism may appear years after graduation, in a certain sense we may say that Phi Beta Kappa itself is a disease. Certainly there is no record of an individual with myxedema attaining Phi Beta Kappa rank."

Despite the assurance of the above observation, it cannot be said that the relationship of thyroid functioning to mental development is as firmly established within the normal range, or among those of superior mental ability, as in the case of specific types of defectives. Ruhberg[15] considers mental retardation to be one of the symptoms of subnormal thyroid function in adults, but this opinion appears to be based solely on clinical observation, and none of his patients were tested under standard conditions.

Since the basal metabolic rate is the best index of the functional activity of the thyroid gland, it is pertinent to consider the correlation, among normal individuals, between *BMR* and intelligence test performance. Hinton[7] has reported surprisingly high coefficients in a group of thirty orphanage children and sixty private-school children (.736 with the Stanford-Binet, .661 with the Arthur Point-Performance Scale). In a later study[8] the same author reports correlations of .70 between *BMR* and Binet IQ and .74 between *BMR* and Performance score for a group of two hundred children aged six to fifteen years. When correlations were calculated for each age separately, with an N of twenty, it was found that the correlation between *BMR* and Binet IQ decreased from .796 $\pm$ .057 at age six to .528 $\pm$ .112 at age fifteen. The correlation between *BMR* and performance decreased from

* Assistance in the preparation of these materials was furnished by the personnel of Works Progress Administration official project numbers 65-3-5406 and 665-08-3-30-Unit A-8.

.765 ± .066 at age six to .483 ± .121 at age fifteen. The greatest drop in the correlation occurred after age ten. It is difficult to evaluate the methods employed in these studies, since the technique in determining the basal metabolic rate is not stated, and no indication is given of the standards which were used for expressing the *BMR*. In a study yielding such strikingly positive results, at variance with earlier studies of normal cases (see footnote on p. 373) it would seem desirable to present a very full account of the experimental procedure, as well as illustrative correlation charts, or tabulations of the original data.

EXPERIMENTAL

Sample: The present report is concerned with a group of approximately forty-three girls and forty-four boys chosen from the population of five elementary schools in Oakland, California. The sample is only fairly representative of the five Oakland schools, some selection having occurred because of the necessity of taking cases regarded as permanent residents, and willing to coöperate in a cumulative test program over a period of seven years.

PHYSIOLOGICAL TESTS

Each child was brought from his home to the laboratory by automobile in the morning before breakfast. After lying supine for a thirty-minute rest period during which repeated pulse rate and blood pressure determinations were made, a Siebe-Gorman half-mask was tied over the child's face with a series of tapes as an insurance against leaks of expired air. After a period of five minutes to permit the child to adjust his breathing through the mask, three determinations of the basal metabolic rate were made by the Tissot open-circuit gasometer method.[13] Eight-minute periods for the collection of expired air were used and analyses of the expired air were made by the Boothby-Sandiford modification of the Haldane technique.[3] Determinations of pulse rate were made by counting over one-minute periods at the end of each eight-minute air collection. At the same time blood-pressure measurements were made by the auscultatory method; systolic pressure was estimated at the first and diastolic pressure at the fourth sound. The basal metabolic rate was computed as heat production in calories per square meter per hour, oxygen consumption in cc. of oxygen consumed per kilogram per minute and also as percentage deviation from the Boothby-Sandiford norms.[1,2] Respira-

tion volume in liters per minute at standard temperature and pressure was calculated from the gasometer volume and temperature readings. Pulse pressures were calculated as the difference between systolic and diastolic pressures. The entire experimental procedure was repeated on the succeeding day for each subject and the average results from the two days' test were used in computing correlations. This testing procedure was carried out on each child at six-month intervals.

In order to minimize the effects of training or practice on the results of the physiological tests, observations collected in the Spring of 1935, when the children had a mean age of fourteen years, were used for the purposes of the present study. This represented the seventh testing of these cases, and it may be assumed that sources of unreliability due to unfamiliarity with the laboratory situation were adequately controlled.

MENTAL TESTS

The intelligence test used was the Terman group test, scores being based upon two administrations (Forms *A* and *B*) at approximately two-week intervals. All of the cases were thoroughly trained in test procedures; the testing program, under the supervision of Dr. Mary Cover Jones, was planned with reference to the maintenance of incentive, and to the control of external factors which might operate to reduce reliability.

RESULTS

Product moment correlation coefficients between mental test scores and the physiological variables were computed for males and females separately and the results are shown in Table I.

The highest correlation obtained is +.27 between Terman group test scores and basal metabolism expressed in calories per square meter per hour. None of the correlations is statistically significant. Examination of the correlation plots failed to indicate any abnormality in the distributions which would tend to lower a correlation coefficient.

It was thought that perhaps the high correlations reported by Hinton were due to the fact that he used results from the first metabolism procedure experienced by his group of children. In the present experiment, habituation to the test procedure was practically complete since each child had been tested seven times previously in the same manner. Hence a similar analysis was carried out, using the physiological data obtained on the first day of the first test, which was

given in the Spring of 1933, when the mean age of the children was twelve years.* The intelligence measures were based on two administrations of the Terman group test in 1933. The correlation coefficients are shown in Table II. None is statistically significant, and many of the coefficients show reversals in direction between the two sexes and between the two experiments for the same variables.

TABLE I.—PHYSICAL-MENTAL RELATIONS IN ADOLESCENTS
44 Males. 43 Females
Mean Age 14.0 Years

	Correlations with mental test		Means		SD	
	Male	Female	Male	Female	Male	Female
Basal metabolism, per cent deviation from Boothby-Sandiford norms	+.07	+.05	− 6.4	− 7.1	7.53	7.21
O_2 consumption, cal./sq. m./hour	+.08	+.27	43.0	38.6	3.49	3.18
Respiration volume, liters per minute	−.07	+.22	5.3	4.8	.70	.54
O_2 consumption, cc./kg./minute	+.10	+.26	4.6	4.0	.46	.42
Basal pulse rate per min	+.07	+.05	68.0	69.4	6.78	7.21
Systolic blood pressure mm. Hg	.00	−.16	111.6	108.9	7.77	5.79
Diastolic blood pressure mm. Hg	−.09	−.11	69.5	68.8	4.42	4.14
Basal pulse pressure mm. Hg	+.07	−.10	43.4	40.7	6.26	5.91
Vital capacity in liters	+.14	+.09	3.2	3.1	.58	.37
Terman group test, mental age	.....	.....	16.7	15.6	1.7	1.6

It is shown elsewhere[16] that the reliability coefficients for the average of six metabolism tests given on two successive days under the experimental conditions outlined is .93. We must, therefore, assume that our average values represent close approximations to the true metabolic rate of the subject tested. Similarly, the reliability of the average of two intelligence tests is of the order of .95. Since we cannot attribute our low coefficients simply to unreliability of measurement, we are forced to conclude that in our group of subjects no relationship exists between intelligence and oxygen consumption recorded under basal conditions. Or if a correlation exists it is too low to be

* Although this is regarded as the first test, all the children had been given a trial period with the apparatus the previous afternoon to familiarize them with the test procedure.

TABLE II.—PHYSICAL-MENTAL RELATIONS IN ADOLESCENTS
46 Boys. 46 Girls
Mean Age 12.0 Years

	Correlations with mental test		Means		SD	
	Male	Female	Male	Female	Male	Female
Basal metabolism, per cent deviation from Boothby-Sandiford norms...	−.15	−.19	− 5.6	− 3.6	7.26	8.55
O_2 consumption, cal./sq. m./hour...	−.11	−.02	44.9	41.6	3.21	3.33
Respiration volume, liters per minute	−.13	+.09	4.6	4.5	.46	.66
O_2 consumption, cc./kg./minute.....	−.01	−.21	5.1	4.6	.49	.56
Basal pulse rate per min.....	+.11	+.09	71.1	73.3	8.10	8.74
Systolic blood pressure mm. Hg.....	−.05	+.21	105.5	106.6	7.75	7.31
Diastolic blood pressure mm. Hg....	−.32	+.03	71.3	73.1	5.13	6.33
Basal pulse pressure mm. Hg.....	+.25	+.19	34.8	34.1	5.83	6.30
Vital capacity in liters.....	No tests made in 1933					
Terman group test, mental age.....	.....	.....	14.3	13.4	1.6	1.3

of any predictive value.* Nor can a relationship be reported between intelligence and basal measurements of blood pressure or pulse rate. It appears that we are here confronted with another example of the

* These findings, although in disagreement with Hinton's report, concur with Patrick and Rowles,[12] who found zero correlations between intelligence and basal metabolic rate and blood pressure in a group of college women. Rothbart,[14] studying boys and girls in a state institution, and Levy,[9] studying clinic children, also found extremely low or zero correlations between intelligence and *BMR*.

Other studies have failed to corroborate the high correlations reported by Hinton. Dispensa[5] reported a correlation of .28 ± .07 between basal metabolic rate and intelligence in a group of seventy-eight mature young women between the ages of seventeen and twenty-three years. In spite of this low correlation she concluded that persons with thyroid function below average may tend to be correspondingly less intelligent and to lack self-restraint, while persons with thyroid functions above average tend to be correspondingly more intelligent, less neurotic, less pessimistic, and less sensitive to worry. The Bernreuter personality schedule and the Humm-Wadsworth test were used in assessing the personality characteristics in the subject. The opposite conclusion was reached by Miles and Miles,[11] who determined the metabolic rates of a group of sixty-five eighth-grade boys by measuring weight loss due to insensible perspiration and evaporation, with a large precision balance. These authors found a lower IQ in the group of children with higher metabolic rate than those with the lower, although the difference was not statistically significant (critical ratio, 2.0).

adaptive mechanism of the human organism. Slight variations in functional activity of the thyroid gland are not reflected in changes in mental capacity because in most individuals other adaptive mechanisms are present which may compensate for this thyroid deficiency. However, as the thyroid deficiency becomes more and more acute a point is reached beyond which compensations cannot occur, with the result that mental retardation appears as a symptom; as the metabolic rate falls from a −15 or −20 to a −40 or −50, mental retardation becomes increasingly manifest, with the appearance of cretinous symptoms. In other words, it may be conjectured that we are dealing with a non-linear function, and are attempting to measure the relationship between two variables with an inadequate technique. Animal experimentation will probably be necessary in testing the above hypothesis.

CONCLUSION

(1) In a group of eighty-seven fourteen-year-old children no correlation was found between measurements of pulse rate or blood pressure taken in the basal state and mental performance on the Terman group test.

(2) No significant correlation was found between basal oxygen consumption, measured as percentage deviation from the Boothby-Sandiford norms, and performance on the Terman group test.

(3) The disagreement between these results and those reported by Hinton raises a question as to whether the differences are due to marked differences in the two samples, or to differences in experimental procedure.

REFERENCES

1. Boothby, W. M., Berkson, J., and Dunn, H. L.: "Studies of the Energy of Metabolism of Normal Individuals: A Standard for Basal Metabolism, with a Nomogram for Clinical Application." *American Journal of Physiology*, Vol. CXVI, No. 2, July, 1936, pp. 468–484.
2. Boothby, W. M., and Sandiford, I.: "Normal Values of Basal or Standard Metabolism. A Modification of the Du Bois Standards." Proceedings of Thirteenth International Physiology Congress, *American Journal of Physiology*, Vol. XC, No. 2, 1929, pp. 290–291.
3. Boothby, W. M., and Sandiford, I.: *Laboratory Manual of the Technique of Basal Metabolic Rate Determination.* Philadelphia: W. B. Saunders Co., 1920.
4. Crile, G.: *The Phenomena of Life.* New York: W. W. Norton & Company, Inc., 1936, pp. 379.

5. Dispensa, Johnetta: "Relationship of the Thyroid with Intelligence and Personality." *The Journal of Psychology*, Vol. VI, July, 1938, pp. 181–186.
6. Gesell, A., Amatruda, C. S., and Culotta, C. S.: "Effect of Thyroid Therapy on Mental and Physical Growth of Cretinous Infants." *American Journal of Diseases of Children*, Vol. LII, November, 1936, pp. 1117–1138.
7. Hinton, R. T.: "The Rôle of the Basal Metabolic Rate in the Intelligence of Ninety Grade-School Students." *Journal of Educational Psychology*, Vol. XXVII, 1936, pp. 546–550.
8. Hinton, R. T.: "A Further Study of the Basal Metabolic Rate in the Intelligence of Children." *Journal of Educational Psychology*, Vol. XXX, 1939, pp. 309–314.
9. Levy, John: "A Quantitative Study of the Relationship Between Basal Metabolic Rate and Children's Behavior Problems." *American Journal of Orthopsychiatry*, Vol. I, April, 1931, pp. 298–310.
10. Lewis, A., Samuel, N., and Galloway, J.: "A Study of Cretinism in London." *Lancet*, Vol. CCXXXII, June 26, 1937, pp. 1505–1509. *Lancet*, Vol. CCXXXIII, July 3, 1937, pp. 5–9.
11. Miles, W. R., and Miles, C. C.: "Personality Type and Metabolism Rate." *Psychological Bulletin*, Vol. XXX, November, 1933, p. 667. (Abstract.)
12. Patrick, James R., and Rowles, Emmett: "Intercorrelations Among Metabolic Rate, Vital Capacity, Blood Pressure, Intelligence, Scholarship, Personality and Other Measures on University Women." *Journal of Applied Psychology*, Vol. XVII, October, 1933, pp. 507–521.
13. Peters, J. P., and Van Slyke, D. D.: *Qualitative Clinical Chemistry—Methods*. Baltimore: Williams and Wilkins Company, Vol. II, 1932, pp. xix–957.
14. Rothbart, H. B.: "Basal Metabolism in Children of Normal and of Sub-normal Intelligence: with Blood Cholesterol and Creatinine Values." *American Journal for Diseases of Children*, Vol. XL, March, 1935, pp. 672–688.
15. Ruhberg, G. N.: "Myxedema, Its Nervous and Mental Manifestations." *Minnesota Medicine*, Vol. XIX, October, 1936, pp. 637–641.
16. Shock, N. W.: "The Reliability of Determinations of Basal Respiratory Functions." (In Preparation.)

A NOTE ON THE INTERCHANGEABILITY OF ART ORIGINALS AND COLORED PHOTOGRAPHIC REPRODUCTIONS

JAMES W. GRIMES

Ohio State University

AND

EDWARD BORDIN

University of Minnesota

INTRODUCTION

The determining of the sensitivity of the observer is a requisite to the study of artistic appreciation. It must be a preliminary to the use of reproductions in the visual arts. Before we can determine what contributes to the esthetic value of an art object, artists' assumptions concerning what they or other observers can distinguish must be tested. An analogous problem arises in the field of music when musicians claim to be able to distinguish between two performances of a composer's work, between the tone of two major orchestras such as the Philadelphia and the New York Philharmonic, or between the recorded and radio concert as compared to a first-hand hearing.

At present most approaches to the testing of art ability and art appreciation have proceeded on the assumption that the abilities to be measured can be adequately sampled by the use of either colored or black and white reproductions. The Meier-Seashore Art Judgment Test—perhaps the most widely accepted art test—uses as test objects black-and-white reproductions of works of old masters paired with alterations of these reproductions. The authors assume that the alteration must necessarily be of less esthetic value than the unchanged reproduction. The McAdory Art Test uses both black-and-white and colored reproductive material. McAdory, however, avoids the problem somewhat by standardizing the test objects by artists' choices of the reproductions. The Christensen Test uses black-and-white reproductions.

The solution of the problem of observer sensitivity is also pertinent to the description of the appreciative process. Too much of the literature on this subject is of a philosophic and discursive nature. More empirical data such as that offered by Maier and

Reninger,[1] relative to problem solving and the creative process, are needed. While Buswell[2] had an interesting approach to the problem in attempting to photograph eye movements of observers, there are questions to be raised about his work from both the technical and the artistic side. There is some question as to the accuracy in plotting the eye movements upon the picture, since charts which he presents must necessarily have multiplied the dimensions of the original record by at least a hundred times. Thus minor discrepancies would become wide errors. Also, there again arises the question of the differential effects that might result from the use of original works as opposed to reproductions and their relation to the art levels of the observers.

On the theoretical side many artists have long contended that a reproduction cannot possibly have the same experiential value for the observer that the original work affords. On the other hand, there have been artists who have conceded even greater value to a colored photographic reproduction than to the original itself. An example of the former point of view is contained in a discussion by Frey[3] of the need for audiences more "sensitive to spiritual values" in creative efforts. His discussion implies that there are very subtle relationships expressed in creative efforts that a reproduction would lose.

The recent technical advances which have been made in the reproductive process have made the answer to this question more pertinent for collectors and art galleries as well as art schools and private individuals. At the outset we wish to emphasize that we are only presenting a tentative attack on the question. We believe that a more decisive answer must await a research in which not one comparison is used, but a large number representing a range in type and quality of work. We hope to stimulate such experimentation under varied conditions through the initiative that our work should offer. It is possible that research with art tests, using proper methodology, might be able to test the adequacy of photographic reproductions by first setting up tests on the basis of originals and, subsequently, studying the effects of a substitution of reproductions of these originals. To our knowledge this has not been done.

---

[1] Maier, N. R. F. and Reninger, H. W.: *A Psychological Approach to Literary Criticism.* New York: D. Appleton & Co., 1933, p. 154.

[2] Buswell, G. T.: *How People Look at Pictures—A Study of Psychology and Perception in Art.* Chicago: University of Chicago Press, 1935, p. 198.

[3] Frey, E. F.: "Artists and Audiences." *Journal of Higher Education,* Vol. IX 1938, pp. 263–270.

PROCEDURE

Through the coöperation of the Columbus Gallery of Fine Arts, the still life, "Yellow Green Chair," by Preston Dickinson, was matched with a collotype reproduction approximately the same size as the original. The reproduction was made by Raymond and Raymond, considered one of the best firms for colored reproductions. The frames on the two pictures being different, a celotex screen was erected with an aperture for each object, permitting only the painting to be seen.

The subject was seated from four to six yards from the apparatus, depending on his vision or inclination. Care was taken to place the apparatus in such a position that the smooth surface of the reproduction would not identify it to the subject by its reflection of light.

The reproduction and original, in that order, were each shown and identified for the subject. They were then presented in random order and the subject asked to indicate whether he though the original or the reproduction was being shown. The subject was warned before hand that the movements behind the screen were standardized and, therefore, would furnish no clues. At any time the subject was given the opportunity to reverse any of his previous identifications. This was to allow for the possibility that a subject might realize he had been consistently identifying the original as the reproduction. The schedule of presentation was as follows: *O, R, O, O, R, O, R, R, R, O.*

At the end of ten trials the subject was questioned concerning the basis on which he had been making his choice. Since most of the subjects wanted to see the two objects together, and since it would also give us information about their distinguishability when presented together, the two objects were shown at the same time and alternated at random until it was evident that the subjects could distinguish them. For some subjects this procedure was not required, for others only one presentation was necessary. Records were kept of any remarks that the subject made during the presentations.

Data were collected for thirty-two subjects. The group was representative of all levels of sophistication in art: Three were members of the Fine Arts faculty at Ohio State University; ten were graduate students in Fine Arts; one was a graduate student in Music; the remaining eighteen were undergraduate students of all four years, for the most part majoring in Fine Arts.

RESULTS

Table I shows us that twenty-two of the subjects were able to make the right designations in eight or more of the ten trials. The composition of this group gives evidence that the experimental situation

TABLE I.—OBSERVED AND THEORETICAL DISTRIBUTIONS OF THIRTY-TWO SUBJECTS ON THE BASIS OF THE NUMBER OF CORRECT CHOICES

Number right of ten	10	9	8	7	6	5	4	3	2	1	0
Theoretical	.03	.32	1.41	3.75	6.56	7.86	6.56	3.75	1.41	.32	.03
Observed	11	6	5	1	2	1	0	2	1	1	2

discriminates differences in training and probably art sensitivity. Two of the three faculty members were included in the ten right group, while the other called eight right. Of the ten graduate students five fell in the ten right, and two in the eight right group.

The first hypothesis made in the treatment of the data was that on a guessing basis there was an equal chance of the presented object being chosen as the original or the reproduction on any trial. Since the subjects did not know how many times in ten trials each object would appear, this assumption seems reasonable.

Table I presents the expected frequency distribution, providing pure chance was operating, compared with the obtained distribution for thirty-two subjects. Because the smallness of the sample gives too many expected frequencies of less than ten, it was felt that the data grouped as in Table II would put the hypothesis to the most rigorous

TABLE II.—DISTRIBUTIONS REGROUPED FOR CHI-SQUARE TEST

Number right of ten	7 to 10	0 to 5
Theoretical	19.93	12.07
Observed	26	6

test. Using the usual formula for Chi-square as given by Snedecor,[1] a value of 4.9 is obtained. This value for one degree of freedom falls between the one per cent and the five per cent points in the Chi-square Distribution[2] reasonably close to the fiducial limit. Thus, it is quite evident that even under the most favorable conditions of calculation the hypothesis of chance is untenable. Without doubt the two objects

[1] Snedecor, G. W.: *Statistical Methods.* Ames, Iowa: Collegiate Press, Inc., 1937, p. 7.

[2] *Ibid.*, p. 163.

caused perceptually different experiences for the majority of the subjects.

The second hypothesis builds upon the foundation laid in the disproof of the first. It is that, although the subjects had different experiences with the two objects, they were only guessing when they identified one experience as being with the original and the other as with the reproduction. If this were so, then, we should expect the obtained distribution to show equal peaks at the extremes. Calculation of Chi-square based on the grouping in Table III gives the significant value of 12.46 for one degree of freedom. Thus we cannot assume that the subjects were guessing in their designations.

TABLE III.—COMPARISONS OF THEORETICAL AND OBSERVED GROUPINGS AT THE EXTREMES OF THE DISTRIBUTIONS

Number right of ten	8 to 10	0 to 2
Theoretical	13	13
Observed	22	4

An examination of the qualitative data contained in the protocols raises some highly suggestive questions which might well form the basis for future research. Eight of the thirty-two subjects indicated some general aspect of the paintings as the basis for their discrimination. All three faculty members were included in this group. Three of the remaining five were graduate students, and the other two were advanced undergraduates. The rest of the subjects indicated that they were differentiating the objects by comparing particular parts such as the color of a bun in the original as compared to the reproduction. One subject was comparing the signature of the artist in the two objects.

The implications that these data may have for some Gestalt interpretation of appreciative behavior must await more critical experimentation. It is also quite possible that these responses are merely verbalizations which are a part of the stereotype of artistic response and have no relation to the real basis on which the objects were discriminated. We are inclined to believe that there is something more than that involved because of some as yet unpublished data which we have on students' memory for paintings.

DISCUSSION

The data presented have shown that the original can be discriminated from the reproduction under the experimental conditions.

Whether this would also be true under the natural conditions in a gallery seems to be dependent on the set of the individual. In a rough preliminary study, the reproduction was substituted for the original on the gallery wall. A few subjects, some of whom were later to take part in the main experiment, were brought in, ostensibly to get their reaction to the painting. Hoping to give them a basis for seeing that there was something different about the object, they were asked to compare it with an adjacent still life also by Dickinson. In all cases there was no definite indication that the subjects were aware that they were comparing a reproduction with an original. But the artistically more sophisticated and, therefore, probably more sensitive subjects could be seen to be responding in terms of their previous experience with the painting in question, rather than the object before them. One subject, a member of the faculty, was observed to be talking about the painting before he had more than glanced at it.

The fact that the colored reproduction is not the equivalent of the original for the entire range of ability of our subjects raises some crucial questions about our existing art tests. For example, could the reproducing of two originals reverse their order of merit? If this could possibly happen in colored reproductions, would it not be even more possible in black-and-white reproductions of colored originals such as have been used in the Meier-Seashore Test? Even if the reversal were not to take place, there would still be a question, in the case of black-and-white reproductions, of whether we are leaving untested an important aspect of the art process—expression through the use of color. This is sufficient to question the validity of existing art-testing instruments.

These questions must be answered before we can accept without reservation the assigning of greater value to one test object rather than another because it is the reproduction of an old master. It is more than likely that the lack of success with existing art tests is directly attributable to these defects. It would seem that, for the present at least, the builders of art tests should avoid techniques involving these questions.

SUMMARY

Thirty-two subjects of a wide range of art sophistication were asked to distinguish an original from its colored reproduction. The results gave indications that, under the experimental conditions, the majority of the subjects could be expected to have visual experiences

with the reproduction which were perceptually different from those with the original. Furthermore, the subjects seem to be able to identify, with a frequency better than chance, which of these different experiences is with the original. This means that the two objects are not behaviorally equivalent.

REFERENCES

Buswell, G. T.: *How People Look at Pictures—A Study of Psychology and Perception in Art.* Chicago: University of Chicago Press, 1935, p. 198.

Frey, E. F.: "Artists and Audiences." *Journal of Higher Education*, Vol. IX, 1938, pp. 263–270.

Karwoski, T. E., and Christensen, E. O.: "A Test for Art Appreciation." *Journal of Educational Psychology*, Vol. XVII, 1926, pp. 187–193.

Kinter, M.: *The Measurement of Artistic Abilities.* New York: Psychological Corporation, 1933, pp. 90.

Maier, N. R. F., and Reninger, H. W.: *A Psychological Approach to Literary Criticism.* New York: D. Appleton & Co., 1933, p. 154.

McAdory, M.: "The Construction and Validation of an Art Test." *Teach. Coll. Contrib. Educ.*, 1929, No. 383, p. 35.

Meier, N. C.: "Measure of Art Talent." *Psychol. Rev. Monog.*, Vol. XXXIX, 1928, pp. 184–199.

Snedecor, G. W.: *Statistical Methods.* Ames Iowa: Collegiate Press, Inc., 1937, p. 341.

A PHOTOGRAPHIC STUDY OF READING DURING A SIXTY-FIVE-HOUR VIGIL

BRANT CLARK

San Jose State College

AND

NEIL WARREN

University of Southern California

INTRODUCTION

Many studies have been conducted by numerous investigators to determine the physiological and psychological reactions during prolonged vigils, but, as far as the authors are aware, very few studies have been made to determine the effect of loss of sleep or fatigue on the behavior of the eyes during reading. The first mention of this subject was made by Dearborn,[8] who found that the accuracy of return sweeps from the end of one line to the beginning of the next was impaired by ocular fatigue. However, he made his observation on but one subject and this finding has not been confirmed.

Miles[5] and Miles and Laslett[6] made studies of horizontal eye movements at the onset of sleep which occurred after a prolonged period of sleeplessness. They found by studying monocular eye-movement records that there was a definite contrast between alertness and drowsiness in the behavior of the eyes. The speed of the saccadic movements was reduced thirty per cent, and the size of the corrective movements was larger, while fixation was very greatly modified. They found that muscular activity seemed to continue for a time after the subject lost consciousness, and that the interfixation movements were reduced to a slow rolling back and forth. Unfortunately, parts of their records were destroyed, and they obtained no photographs of persons falling asleep while they were reading.

Kurtz[4] in an experimental investigation of ocular fatigue studied the eye movements of six subjects before and after a thirty-minute period of intense visual effort. It was found that reading speed decreased from five to sixteen per cent, and the number of fixations decreased slightly for one subject studied. No detailed report was given of the change in fixations, regressions, and binocular behavior. The significance of the data is limited by the small number of subjects, and the fact that only beginning and terminal tests were made,

although it is significant that all subjects showed a decrement in reading speed.

It was the purpose of the investigation reported here to make a detailed and periodic study of eye movements during reading throughout a sixty-five-hour vigil and to determine any changes in fixations, regressions, and binocular behavior of the eyes.

EXPERIMENTAL PROCEDURE

Since the experimental procedure has been described in detail elsewhere[3,9] only a brief description will be included here. Four psychology students at the University of Southern California acted as subjects during the vigil, which lasted approximately sixty-five hours. A preliminary test was given prior to the vigil and a post-test was given after the subjects had had at least one night of sleep. Experimenters and assistants worked in shifts, and the tests were made every ten hours on a previously determined schedule. Photographic records of reading were only one of a series of tests given. Partial results have been reported elsewhere.[3,9] These tests kept the subjects occupied a great part of the time and for the remainder of the vigil they were kept awake by walks and games of various sorts.

The reading records were obtained by a binocular eye-movement camera.[2] The subjects read short selections from material which was relatively easy for all of them. At the time of the experiment, the experimenters were chiefly concerned with binocular factors in reading, so, unfortunately, comprehension scores were not systematically recorded. However, the subjects were instructed to read so that they could give an accurate summary of the material and in all cases they were able to do this, giving details which indicated satisfactory comprehension.

Eye-movement records were obtained every ten hours during the vigil, and one additional test was made at forty-six hours. All of the records were legible, and data were obtained throughout the sixty-five hours for three subjects, and through forty-six hours for one subject who was unable to continue beyond that time.

RESULTS

All of the photographic data were treated by skilled assistants* who had had considerable experience in the measurement and tabula-

* The assistance of the San Jose State College WPA Project No. 8284 in obtaining the data from the photographic records is gratefully acknowledged.

tion of eye-movement records. The records were measured directly by means of a millimeter rule, the measurements being made to the nearest one-half mm. One mm displacement on the record indicated an ocular rotation of forty-eight minutes. Time measurements were in terms of one twenty-fifth seconds. All records were measured at least twice and in certain cases one of the authors made further checks which indicated a high reliability of the measurements. The fixations, regressions, and fixation time were counted, and binocular adjustments during the fixation pauses were determined. Binocular adjustments were measured for over three thousand fixation movements during reading.

Fixations and Regressions.—The data for all four subjects are summarized in Tables I and II. These are the means for the four subjects, but the behavior of each *S* approximated the average. Preliminary tests are not included because they were not given to all subjects. An examination of Table I indicates that there was no uniform change in the number of fixations and regressions per line, nor in the average fixation time. Although the greatest number of fixations per line was found in the final test, there was no consistent change up to that point. The greatest deviation in regressions was the small number found at forty-six hours.

TABLE I.—AVERAGE FIXATIONS, REGRESSIONS, AND FIXATION TIME DURING READING THROUGHOUT A SIXTY-FIVE-HOUR VIGIL

Hours of sleeplessness.....	0	10	20	30	40	46	50	60	Post-
Average fixation per line...	10.69	10.97	7.92	10.00	9.89	8.32	10.38	12.60	8.85
Average regression per line.	2.06	2.16	1.93	1.79	1.72	0.70	1.62	1.82	1.54
Average fixation time in $\frac{1}{25}$ second.....	5.30	5.60	6.10	4.90	5.20	5.30	4.80	5.30	5.00

A study of the data tabulated for each subject showed greater individual variation from one test to another, but the changes that did occur were definitely sporadic. Not one of the subjects exhibited a uniform change in any measurement made. This was particularly true of the average fixation time, which showed very little variation. The maximum range of variation throughout the series was two twenty-fifths seconds for subject *St*, while one subject (*Fr*) showed a range of average scores which was only one-fiftieth second. The most marked variation in behavior occurred with a subject (*St*) who fell asleep while he was reading. These variations will be discussed in detail under binocular adjustments, but it is important to note that

two or three minutes after he was awakened, his ocular behavior as shown by photography was that typical of his "normal" reading.

Binocular Adjustments.—During reading, the eyes make certain binocular adjustments, *e.g.*, during the initial fixation the eyes diverge, and they also diverge as the subject reads through the line. The divergent movements were particularly marked with these subjects because of the long lines used (fifteen cm). The data for three of the binocular adjustments measured are summarized in Table II. The measurements for the first fixation were taken by determining the position of the eyes at the initial and final parts of the fixation, and the time required to complete the divergent adjustment was measured by determining the point at which the divergent movement ceased. The divergence from the first to the final fixation was measured by determining the position of the eyes at the beginning of the first fixation and comparing it with their position at the end of the final fixation.

An examination of Table II indicates no uniform variation in the extent of the binocular adjustment. Considerable variability was noted throughout, but this is common in subjects who are not fatigued. However, the time required to complete the divergent movements at the beginning of the lines was considerably longer on the final tests.

TABLE II.—AVERAGE BINOCULAR ADJUSTMENTS DURING READING THROUGHOUT A SIXTY-FIVE-HOUR VIGIL

	Hours of sleeplessness								
Average divergence.....	0	10	20	30	40	46	50	60	Post-
First fixation in minutes.....	56.2	58.1	33.6	40.3	34.6	18.7	48.5	44.6	64.3
First to final fixation in minutes.....	103.8	64.3	56.6	63.4	41.8	68.2	111.4	118.1	127.7
Time for first fixation in $\frac{1}{25}$ second.....	2.59	1.83	1.86	2.42	1.46	1.21	3.02	3.41	2.78

An interesting record was that obtained from a subject (*St*) who fell asleep while he was reading in front of the camera. It is particularly interesting because such records can hardly be taken by prearrangement, and catching a subject just at the point of falling asleep is a matter of rare chance. Miles[5] obtained such a record during simple interfixation movements, but was not fortunate enough to get one during reading. In this case the subject fell asleep during the return sweep to the beginning of the second line.

The record showed marked deviations in ocular behavior, not only in irregular fixations, but in binocular behavior. For this material the subject in normal reading averaged approximately eleven fixations per line, but in the line which he "read" just prior to falling asleep the eyes made but four typical saccadic movements. The remainder of the distance was executed by a slow, more or less continuous movement along the line. The first saccadic movement was a regression at the beginning of the line, which occurred normally in a high per cent of the cases. This was followed immediately by two saccadic movements and then the eyes moved slowly across the line. The eyes blinked about two-thirds of the way across and made a saccadic movement at the end of the line. During the blink the eyes moved up and diverged as the lid came down, then converged and moved down as the lids opened. There were no outstanding vertical irregularities. The "reading time" for the line was three and twenty twenty-fifths seconds, while the average was approximately two and one-half seconds.

The binocular adjustments showed marked deviations from the usual pattern. The eyes diverged 2.8° from the beginning to the end of the line, while the subject's average was 1.1° and his next largest divergence was 1.9°. The greatest binocular disparities occurred in making the movement from the end of the first line to the beginning of the second. The right eye made the excursion in eighteen twenty-fifths seconds, wavering back and forth during fixation for eighteen twenty-fifths seconds just before the eyes closed. The left eye lagged behind the right eye and arrived at the fixation point in twenty-nine twenty-fifths seconds just as the left eye closed. The speed of movement decreased as the eye approached the fixation point. The right eye remained open seven-twenty-fifths seconds longer than the left. During the latter part of this horizontal movement the eyes also were slowly moving up.

Speed of Movement.—The unit of time measured with this camera was one-twenty-fifth second, so that time could be estimated to one-fiftieth second. This was too gross a measure for an accurate comparison of the time required to complete the return sweeps to the beginning of the lines. However, the time required was about three-fiftieths seconds, which approximates Tinker's[7] more accurate results.

One subject showed no measureable variation in speed of movement throughout the vigil, whereas the other three subjects showed increases in time required to complete the return sweeps to the begin-

ning of the lines. These decreases in speed of movement all occurred after forty hours of wakefulness. The speed of movement decreased the most for *St*, whose records have been discussed above. However, the speed of movement returned to "normal" in a record taken just after he was awakened and the average time to complete the return sweeps for the two final tests was approximately two and one half twenty-fifths seconds. *Th*'s average time for the return sweeps at the forty-hour test was three and four tenths twenty-fifths seconds, at the fifty-hour test, two and six tenths twenty-fifths seconds; and at the final test two and three tenths twenty-fifths seconds. At the forty-hour test *Li*'s average was two and nine tenths twenty-fifths seconds and at the forty-six-hour test (his last series) it was four and one half twenty-fifths seconds.

These data indicate a tendency toward decrement in the speed of ocular movement during the return sweeps to the beginning of the lines. However, it is significant to note that these changes, like others reported, were definitely sporadic and were probably indications of the presence of blocking, as has been reported by Bills.[1] No attempt was made to measure speed of movement between fixations within the line because of the grossness of the time measurement.

Table II indicates that there was also no regular change in the time required to complete the divergent movements at the beginnings of the lines. However, the final two tests showed an increase for all subjects.

SUMMARY AND CONCLUSIONS

An analysis of binocular photographic reading records taken during a prolonged vigil showed no uniform changes in the behavior of the eyes as the period of wakefulness increased. Although there were certain marked changes in the behavior of individual subjects, in general these changes were uniformly temporary and sporadic, rather than regular. They are comparable to the findings of "blockings" during alternate addition and subtraction, and color-naming for the same subjects.[9] An increase in the number of "blocks" occurred in these activities after approximately forty hours of wakefulness. The eye-movement records showed similar periods late in the vigil during which there was a marked reduction in the ability to perform adequate fixation movements.

The fixation time, fixations, and regressions showed no consistent variation. The magnitude of binocular adjustments showed no uni-

form trend although there was a tendency toward decrement of the time required to complete the divergent movements at the beginning of the lines.

The speed of the saccadic movements to the beginning of the lines decreased sporadically for three of the subjects. The time required to complete the movement was as much as two and one-fourth times the average for *Th*, three times the average for *Li*, and nineteen times the average for *St* just before he fell asleep. This reduction in speed of movement was far from consistent, however. In two of the subjects there was at least a partial recovery on subsequent tests.

The greatest deviations from normal occurred in one subject who fell asleep during the test. Saccadic movements were replaced by a slow, continuous movement across the page. As the eyes moved to the next line, the speed of movement showed a marked decrease, and the left eye lagged behind the right, making a marked binocular disparity. A low level of awareness was indicated until just before the lids closed by the fact that the eyes moved toward the beginning of the next line.

It is clear that the vigil was not prolonged to the point of exhaustion. The changes which did occur may be explained as a temporary failure to overcome the greater subjective threshold of attention and effort.[10] These "blocks" or periods of reduction in ability to respond give brief rest periods which permit the subject to continue to respond on subsequent performances.

The data presented also show the importance of periodic tests during experimental fatigue. Initial and terminal tests alone do not give an adequate picture of the situation. The data also show the importance of designing tests of fatigue which make compensation either difficult or impossible, and that even objective studies of eye movements which are largely involuntary fall short of adequately achieving this purpose.

BIBLIOGRAPHY

1. Bills, A. G.: "Fatigue in mental work." *Physiol. Rev.*, 1937, Vol. xvii, pp. 436–453.
2. Clark, B.: "A camera for simultaneous record of the horizontal and vertical movements of both eyes." *Amer. J. Psychol.*, 1934, Vol. xlvi, pp. 325–326.
3. Clark, B. and Warren, N.: "The effect of loss of sleep on visual tests." *Am. J. Optom.*, 1939, Vol. xvi, pp. 80–95.
4. Kurtz, J. I.: "An experimental study of ocular fatigue." *Am. J. Optom.*, 1938, Vol. xv, pp. 86–117.

5. MILES, W. R.: "Horizontal eye movements at the onset of sleep." *Psychol. Rev.*, 1929, Vol. XXXVI, pp. 122–141.
6. MILES, W. R. AND LASLETT, H. R.: "Visual fixation during profound sleepiness." *Psychol. Rev.*, 1931, Vol. XXXVIII, pp. 1–13.
7. TINKER, M. A.: "Eye-movement duration, pause duration, and reading time." *Psychol. Rev.*, 1928, Vol. XXXV, pp. 385–397.
8. VERNON, M. D.: *The experimental study of reading.* Cambridge (Eng.): The University Press, 1931.
9. WARREN, N. AND CLARK, B.: "Blocking in mental and motor tasks during a sixty-five-hour vigil." *J. exp. Psychol.*, 1937, Vol. XXI, pp. 97–105.
10. WHITING, H. F. AND ENGLISH, H. B.: "Fatigue tests and incentives." *J. exp. Psychol.*, 1925, Vol. VIII, pp. 33–49.

NOTE ON THE CORRELATION OF INITIAL SCORES WITH GAINS

LESLIE ZIEVE*

University of Minnesota

In psychological work, the problem of obtaining a measure of relationship between initial scores and gains frequently arises, with the result that a coefficient of correlation between the variates is computed. However, use of the ordinary product-moment formula results in a spurious relationship that requires correction. The spuriousness, which exists as a reduction in the correlation below its true value, is produced by the presence of an inverse relationship between the errors in the observed initial and gain scores. The error of the initial score is present in the gain score with the same magnitude but with opposite sign. How this comes about is most explicitly seen in the algebraic statement of the problem.

Suppose we let

x = observed initial score
a = true initial score
e = error in initial score
y = observed gain
g = true gain
z = observed final score
e' = error in final score
r_x = coefficient of reliability for x scores
r_y = coefficient of reliability for y scores
r_z = coefficient of reliability for z scores

Then

$$\begin{aligned} x &= a + e \\ z &= a + g + e' \\ (z - x) = y &= g + (e' - e) \end{aligned} \qquad (1)$$

where, without loss of generality, we consider all scores as deviations from their respective means. In the first and third of these equations e is present as a positive and negative (algebraically) quantity, respectively. It is because of this that r_{xy} gives a spurious value for the correlation between initial scores and gains.† What is required is r_{ag},

* From the Institute of Child Welfare.

† In an article by Thorndike[3] will be found detailed numerical examples illustrating this point.

the correlation between true initial scores and true gains. With the set-up as given, and assuming that errors e and e' are uncorrelated with either a or g or with each other, G. H. Thomson[1] first derived the correct formula

$$r_{ag} = \frac{r_{xy} + \frac{\sigma_x}{\sigma_y}(1 - r_x)}{\frac{1}{\sigma_y}\sqrt{r_x}\sqrt{\sigma_y^2 - \sigma_x^2(1 - r_x) - \sigma_z^2(1 - r_z)}}; \tag{2}$$

and from this he[2] developed the alternative formula

$$r_{ag} = \frac{\sigma_z r_{xz} - \sigma_x r_x}{\sqrt{r_x}\sqrt{r_x\sigma_x^2 + r_z\sigma_z^2 - 2\sigma_x\sigma_z r_{xz}}}. \tag{3}$$

It is possible to present a simplification which, though obvious enough to anyone who looks into the problem, deserves explicit mention because it gives an easily remembered formula and reduces the labor of computation materially. Assuming in equations (1) that of a, g, e, e' only a and g may be correlated, multiplying $x = a + e$ by $y = g + e' - e$, and summing, we get $\Sigma xy = \Sigma ag - \Sigma e^2$ which when divided by the number of cases gives the equation

$$r_{xy}\sigma_x\sigma_y + \sigma_e^2 = r_{ag}\sigma_a\sigma_g. \tag{4}$$

It is necessary to determine the values σ_e^2, σ_a, σ_g; and these may be gotten from the relations, $\sigma_x^2 = \sigma_a^2 + \sigma_e^2$, $\sigma_a^2 = \sigma_x^2 r_x$ [see 4, p. 298], $\sigma_g^2 = \sigma_y^2 r_y$. It is seen that $\sigma_e^2 = \sigma_x^2(1 - r_x)$, $\sigma_a = \sigma_x\sqrt{r_x}$ and $\sigma_g = \sigma_y\sqrt{r_y}$. Substituting in (4) and rearranging, we get the simplified formula

$$r_{ag} = \frac{r_{xy} + \frac{\sigma_x}{\sigma_y}(1 - r_x)}{\sqrt{r_x}\sqrt{r_y}}. \tag{5}$$

That formula (5) is identical with (2) can be demonstrated by verifying the identity, $\sqrt{r_y} = \frac{1}{\sigma_y}\sqrt{\sigma_y^2 - \sigma_x^2(1 - r_x) - \sigma_z^2(1 - r_z)}$, and this is easily done when use is made of the equations, $\sigma_y^2 = \sigma_g^2 + (\sigma_e^2 + \sigma_{e'}^2)$, $\sigma_z^2 = \sigma_{a+g}^2 + \sigma_{e'}^2$, and $\sigma_{a+g}^2 = \sigma_z^2 r_z$, together with those already given.

The form of (5) brings out its close resemblance to the ordinary correction for attenuation formula

$$\frac{r_{xy}}{\sqrt{r_x}\sqrt{r_y}}, \tag{6}$$

the only difference being the additional term in the numerator of (5). The principles behind the derivations of (5) and (6) are identical, but the formulas are slightly different because in the former the error in y is dependent upon that in x, whereas in the latter the error in y is independent of that in x. In the special case where $r_x = 1$ both (5) and (6) reduce to the same quantity, *e.g.*, $\frac{r_{xy}}{\sqrt{r_y}}$. When x and y are two measures whose errors are independent, attenuation in the observed correlation becomes increasingly large as the reliabilities of the measures decrease. Specifically, r_{xy} is the fraction of r_{ag} corresponding to the value, $\sqrt{r_x r_y}$; if $r_x = r_y$ the fraction is given by r_x. Thus if $r_{ag} = .90$ and $r_x = r_y = .80$, then r_{xy} is $\frac{4}{5}$ of .90 or .72. The attenuation effect produced when initial and gain scores are correlated is even more pronounced, and a sizeable error in interpretation is incurred if the correlation is not corrected. To illustrate: If the true correlation, r_{ag}, were zero, we would have $r_{xy} = \frac{\sigma_x}{\sigma_y}(r_x - 1)$ which is negative when r_x is not perfect (as is usual). A negative relationship is inferred when actually there is none. This point is illustrated numerically in the first column of Table I, in which are tabulated the values of r_{xy} corresponding to various values of r_{ag} and $r_x = r_y$ (assuming $\frac{\sigma_x}{\sigma_y} = 1$). It is possible to have a materially positive value for r_{ag} and still get a

TABLE I.—VALUES OF r_{xy} CORRESPONDING TO VARIOUS VALUES OF r_{ag} AND $r_x = r_y$
(TAKING $\sigma_x/\sigma_y = 1$)
(Entries Calculated by Formula 5)

Values of $r_x = r_y$	Values of r_{ag}										
	0	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
0	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00
.10	− .90	− .89	− .88	− .87	− .86	− .85	− .84	− .83	− .82	− .81	− .80
.20	− .80	− .78	− .76	− .74	− .72	− .70	− .68	− .66	− .64	− .62	− .60
.30	− .70	− .67	− .64	− .61	− .58	− .55	− .52	− .49	− .46	− .43	− .40
.40	− .60	− .56	− .52	− .48	− .44	− .40	− .36	− .32	− .28	− .24	− .20
.50	− .50	− .45	− .40	− .35	− .30	− .25	− .20	− .15	− .10	− .05	0
.60	− .40	− .34	− .28	− .22	− .16	− .10	− .04	+ .02	+ .08	+ .14	+ .20
.70	− .30	− .23	− .16	− .09	− .02	+ .05	+ .12	+ .19	+ .26	+ .33	+ .40
.80	− .20	− .12	− .04	+ .04	+ .12	+ .20	+ .28	+ .36	+ .44	+ .52	+ .60
.90	− .10	− .01	+ .08	+ .17	+ .26	+ .35	+ .44	+ .53	+ .62	+ .71	+ .80
1.00	0	+ .10	+ .20	+ .30	+ .40	+ .50	+ .60	+ .70	+ .80	+ .90	+1.00

negative value for r_{xy}. Such is the case for $r_{ag} = .50$, $\frac{\sigma_x}{\sigma_y} = 1$, and $r_x = r_y = .60$, r_{xy} having the value $-.10$ which is given in the seventh row and sixth column of the table. If we merely specify that r_{ag} be positive, the lower bound for r_{xy} is $\frac{\sigma_x}{\sigma_y}(r_x - 1)$, and this permits substantial negative values for r_{xy}.

REFERENCES

1. Thomson, G. H.: "A formula to correct for the effect of errors of measurement on the correlation of initial values with gains." *J. Exp. Psych.*, Vol. VII, 1924, pp. 321–324.
2. ———: "An alternative formula for the true correlation of initial values with gains." *J. Exp. Psych.*, Vol. VIII, 1925, pp. 323–324.
3. Thorndike, E. L.: "The influence of the chance imperfections of measures upon the relation of initial score to gain or loss." *J. Exp. Psych.*, Vol. VII, 1924, pp. 225–232.
4. Yule, G. Udney and Kendall, M. G.: *An Introduction to the Theory of Statistics.* Charles Griffen and Company, Limited, London, 1937.

BOOK REVIEWS

ALBERT J. HARRIS. *How to Increase Reading Ability.* New York: Longmans, Green and Co., 1940, pp. 403.

The attitude assumed by this author toward reading disabilities is most sensible. He contends that the great majority of remedial cases arise "from relatively simple causes" such as mental or social immaturity, sensory handicaps, poor motivation, absence from school, poor teaching, and so on (p. 19). The "special mental defect" hypotheses are properly minimized thusly: "The terms 'congenital word blindness' and 'development alexia' . . . (have) gradually died out as greater success has been attained in diagnosis" (p. 136).

Harris describes his text as being a practical one, and the laudable desire to be helpful gets him into difficulties now and then. For example, even though he recognizes the inadvisability of any specific time allocation to oral as distinguished from silent reading he—because "some answer must be made to the inquiring teacher"—recommends devoting more than half of the instruction period to oral reading during the first year and less than half during the second and third years. After this, attention to silent reading should increase until two-thirds of the time of fourth-graders and not less than three-fourths of the time of fifth-graders is so occupied (p. 43f). Despite this specificity, Harris would in all likelihood admit that the relative amount of attention to be devoted to oral or silent reading should depend primarily if not entirely upon pupil progress and an analysis of reading difficulties.

The treatment of remedial reading *per se* is confined to the last three-fifths of the book and is in general excellent. The preceding one hundred seventy pages include a comprehensive discussion of reading—its nature, methods of teaching in general, a stimulating chapter on reading readiness, and some forty pages dealing with the cause of reading difficulties. The chapter on reading readiness includes much material on testing for readiness although, strangely enough, there is no emphasis upon clinical evidence for readiness such as the tendency for some children to seek reading experiences. The treatment of diagnosis (silent and oral) contains a complete description of available tests. Chapter VI, "Investigating the Causes of Reading Difficulties," describes means for getting at the following factors related to reading disability: (*a*) Intelligence, (*b*) special mental

defects, (*c*) visual and auditory acuity, (*d*) muscular coördination, (*e*) illness, (*f*) glandular disturbances, (*g*) hand-eye dominance, (*h*) the school record, and (*i*) personality and home background.

The reviewer was slightly disappointed in the chapters dealing with specific remedial materials and practices. He knows as much about the field as does the typical classroom teacher who might be expected to benefit most from the text, and he felt the need for something more explicit. Long case studies would have helped—lengthy descriptions of just what was done (with success) from the time a particular child was suspected of reading limitations until he no longer received special treatment. There are only three of these longer descriptive passages (p. 186–200) and they are included primarily to illustrate the case-study technique. This criticism seems all the more pertinent in view of the editor's introductory statement that the text is a treatment of *how to* increase reading ability. In the reviewer's judgment it is more a good summary of how reading ability *has been* increased. The difference between these aims is more real than apparent.

The research literature referred to by Harris is thoroughly up to date and even the latest citations are an integral part of the book—not added to the page proof. Chapter VI, "Investigating the Causes of Reading Difficulties," is buttressed by more than forty well-selected research references of recent date. In view of the "practical" emphasis throughout the text, this gives the critical reader much more confidence in his author than he might otherwise have. The format of the book is superior. Appendix A contains a complete alphabetical list of all the tests described, and Appendix B a helpful graded list of books for supplementary reading. The index is sufficiently complete to be useful.

STEPHEN M. COREY.
University of Wisconsin.

LAWRENCE E. COLE. *General Psychology.* New York: McGraw-Hill Book Co., 1939, pp. 688.

What are the desirable qualities in an elementary textbook? The answer to this question will vary in its details with each instructor of an elementary class. However, all would agree that a simple style of writing, insightful selection from the experimental literature, a consistent systematic position with due regard to differing points of view, and an integration of experimental results with human life are all necessary. Cole's book has all of these in full measure.

An historical chapter, "Animism and Brain Psychology," discloses the sources of many popular fallacies concerning the mind and the functions of the nervous system. In the next three chapters the psychological significance of the nervous system, receptors and effectors are described. The subjects of the remaining chapters are development, emotion, motivation, learning, perception, thinking, reasoning, intelligence, and personality. While this organization may appear peculiar to the conventionally-minded, the reader will find a definite coherence in the sequence.

The author disclaims any attempt to write within the rigid confines of a system, but he warns that he has a behavioristic bias. However, he promises in the preface that he shall attempt to present the viewpoints of other schools, such as Gestalt and Freudian psychoanalysis, on controversial points. This he does fairly, and in such a manner to avoid confusing the student. His behaviorism is tempered. Perhaps the most succinct statement of the book's tenor is in a quotation from the last paragraph: "No longer can we confine our observation to a peering within the biological organism or to an introspective analysis of individual consciousness. While we might say that our science is rooted in biology, we must see to it that it is oriented toward the social."

I have read most of the elementary textbooks published during the past decade, and have reviewed a number of them. Cole's book is without doubt deserving a place in the highest semi-decile. It should take equal rank in its usefulness with those classics of the elementary field, Woodworth and Dashiell.

C. M. Louttit.
Indiana University.

Alan E. Treloar. *Elements of Statistical Reasoning.* New York: John Wiley and Sons, 1939, pp. 261.

It has been a real pleasure to find a statistical text which combines simplicity and conciseness and avoids being superficial. This book, by a student of the late J. Arthur Harris, treats of the usual topics found in texts in this field: Frequency distributions and their description, correlation, sampling, and attributes in terms of proportion and Chi-Square. The treatment of correlation is particularly penetrating. Partial and multiple correlation are not included; nor are there exercises. Aside from a chapter on vital statistics, psychologists will find that Treloar has discussed many of the techniques needed in

psychological research. For classroom use, the reviewer would prefer to have the discussion of probability precede that of sampling, but this does not detract from the merits of a soundly-written book.

QUINN MCNEMAR.

Stanford University.

JAY W. FAY. *American Psychology before William James.* New Brunswick: Rutgers University Press, 1939, pp. 240.

The notion that no psychology worthy of the name existed in America prior to 1880 has been fostered both by disparaging remarks and by neglect of the period by historians of psychology. Fay's survey of textbooks used from 1640 to 1890 vigorously refutes this view. He claims that such a view is no truer than the statement that there was no psychology in Europe before Wundt. He admits, however, that the earlier psychology was quite different in pattern from the later experimental or scientific psychology. The pre-James psychology falls rather naturally into three eras: Period of theology and moral philosophy, 1640–1776; period of intellectual philosophy, 1776–1861; and period of British and German influences, 1861–1890. Although the early part of the first period was one of intellectual isolation, the latter part reveals some influence from European philosophy, especially British. The second period was dominated by Scottish philosophy. In the third period there was an increasing dependence upon British associationism, some signs of accepting the evolution theory and obvious effects from German psychophysics and physiological psychology. These became prominent during the ten years just prior to the publication of James' *Principles.*

Early American psychology, of course, was within the discipline of philosophy. Psychological trends are related in an interesting manner to the conditions of the country. To a surprising extent one finds that topics which were to bulk large in American psychology of the Twentieth Century found a place in the early texts. Included were: Genetic and child, individual, comparative, social, and *abnormal* psychology, religious experience, *mental hygiene,* instinct, habit, personality, and educational applications.

At times Fay becomes overly enthusiastic concerning the relative importance of certain contributions: Johnson's text (1752) is considered the equal of anything the other side of the Atlantic. Upham is called the Sully of America. Few will agree that Johnson, Smith, Rush,

Burton, Upham and Hickok "made as real contributions to the science of psychology as James, Baldwin, Watson, and Thorndike have since made to the science which has taken its place, while still bearing its name."

While one must admire the thoroughness of Fay's survey he apparently missed the highly important contribution of Joseph Buchanan's *Philosophy of Human Nature*, 1812, which presents a sound biological, deterministic viewpoint. Furthermore, the influence of orthodoxy in determining the contents of many Nineteenth Century texts might have been emphasized. The author's method of quoting frequently and at length becomes decidedly tedious to the reader.

Fay is to be commended for producing a comprehensive treatise on a period of American psychology with which few are acquainted.

MILES A. TINKER.

University of Minnesota.

N. L. MUNN. *Psychological Development.* Boston: Houghton Mifflin Company, 1938, pp. 582.

Psychological Development fills a much felt need for a solid, up-to-date text in genetic psychology. Lacking the discursiveness of the usual text in child psychology, it handles a specially selected group of problems—problems which for the most part are of systematic importance and are defined by an interest in a fuller understanding of adult phenomena. Experimental phylogenetic evidence is introduced to parallel the ontogenetic material and comprises approximately one-fourth of the book. Arranged in a well-ordered sequence, the major topics covered include: The mechanisms of inheritance, the heredity-environment problem, unlearned behavior in animals, the evolution of the receptor mechanisms and the nervous system, the development of learning ability in animals, embryological development and the sequence of behavior before birth, maturation and learning in infant behavior, the development of receptor function in infants, the development of space perception, motor coördination, memory and thought, intelligence, emotional and social behavior, and, in conclusion, the changes in personality from adolescence to senescence.

The strength of the book lies in its experimental and physiological approach. Munn reviews the literature thoroughly, as indicated by his bibliography which extends over thirty-five pages, and on controversial issues offers a sympathetic evaluation of opposing points of

view. As dictated by this comprehensive survey of experimental results, the book is tersely written and conservative in its conclusions. The advanced student finds this method of presentation particularly stimulating, but the beginner, unless well oriented, often comments that the discussions are hurried and indefinite. All agree, however, in admiring the organization of the book and its wealth of factual material.

The complete exclusion of topics of clinical and typically popular interest, delinquency and conduct problems, hygiene of development and the like, would make it appear that *Psychological Development* might be limited or specialized in its application, but it should prove valuable in second year courses in advanced general psychology and as a coördinate text in child and educational psychology.

WILLIAM E. KAPPAUF.

University of Rochester.

EMILY L. STOGDILL AND AUDELL HERNDON. *Objective Personality Study.* New York: Longmans, Green and Co., 1939, pp. 106.

This is a work-book designed for college courses in mental hygiene in which the emphasis is upon the student's own adjustment. The sheets are perforated so that they may be filled out and handed in at various times. A great variety of information is called for with regard to the student's life history, emotional adjustments, and relationships to other persons. The work-book thus takes the place of the autobiography required in many mental hygiene courses, and gives the student occasions for making a frank, objective appraisal of himself.

The booklet appears to the reviewer to be an excellent device for use in such courses.

MELVIN G. RIGG.

Oklahoma A. and M. College.